STEPS TOWARDS THE DEVELOPMENT OF AN AUTOMATIC CLASSIFIER FOR ASTRONOMICAL SOURCES

Carole Thiebaut, Michel Boër, Mathieu Bringer
CESR-CNRS
9 avenue du Colonel Roche, BP 4346, 31028 Toulouse Cedex 4
Tel : 33 561 556 641; Fax: 33 561 556 701
e-mail: carole.thiebaut@cesr.fr, michel.boer@cesr.fr

ABSTRACT

We present the results obtained in implementing an automatic classifier for astronomical objects. We studied different neural network architectures for the classification of object found in astronomical images (2D case) and we are now implementing a classifier which works both in the image (2D) and time domain. The 2D classifier is based on a Self Organizing Map. The method we describe is adaptive, is trained by examples and doesn't need any training rules. The map is used after training with TAROT objects (Télescope à Action Rapide pour les Objets Transitoires, *Rapid Action Telescope for Transient Objects*).

In this paper, we present the classifiers we tested, and we describe our 2D classifier method as well as the results from simulated and real astronomical images. We present also the next step of classification through our 3D (geometry – time) classifier. In general our method works better than other automatic methods, but needs that an extensive set of all kind of sources, including those rarely encountered, is presented in the training set.

1. INTRODUCTION

The primary goal of the Télescope à Action Rapide pour les Objets Transitoires (*Rapid Action Telescope for Transient Objects, hereafter TAROT;* Boër et al. 1999, Boër et al. 2001) is the simultaneous observation and rapid detection of cosmic Gamma-Ray Bursts (hereafter GRBs) at gamma-ray and visible wavelengths. We have developed an automatic processing software system, which we are now linking with a classifier. For the moment, the classification is based on the geometrical characteristics of the object (the so-called 2D classifier). But in order to recognize optical counterparts of cosmic GRBs and more generally to classify the variable objects, we have to take into account the temporal profiles of the sources. We have then to develop a classifier which will be based both on geometrical and temporal characteristics of the objects (the so-called 3D classifier).

This paper deals mainly with the so-called 2D classifier, based on the geometrical characteristics of the objects. We present however the progress we have made in taking into account the temporal properties of the sources. The next section describes the TAROT autonomous observatory. In section 3 we present the 2D classifier and the results obtained using it. The last section is devoted to the conclusion and perspectives of this work, including a discussion on the 3D classifier.

2. THE TAROT AUTONOMOUS OBSERVATORY

TAROT is a fully automated 25 cm aperture telescope (Bringer et al. 2001). Table 1 summarized the current main technical characteristics of TAROT.

Aperture	25 cm
Field of view	2deg x 2deg
Optical resolution	20 µm
Mount type	Equatorial
Axis speed	Adjustable, up to 80deg/s
CCD type	Thomson THX 7899
CCD size	2082 x 2072 pixels
Pixel size	15 µm
CCD readout noise	~ 14 e-
Readout time	2 s
Filter wheel	Clear, V, R, I, B+V, R+I

Table 1: Main technical characteristics of TAROT.

As soon as a frame has been acquired, it is processed through the data processing pipeline which produces an output catalogue containing the extracted sources and their characteristics (geometry, radiometry…). The image processing is made of different steps: the first one is the removal of bias, dark and flat field calibration frames. The second one is the computation of a background estimation image and its subtraction to the original one. Then, both steps of source detection and separation are provided by the Lutz's algorithm (Lutz 1979). The final step is the measurements of object characteristics.

Stars, galaxies, plane and satellite tracks are detected by our software. The results of the astrometrical and photometrical calibrations with the USNO-A.2.0 catalogue are included in the output catalogue as well as the object position, the magnitude, the flux, and the number of pixels composing the object.

Moreover, the software is able to produce sub-images of sources that are not larger than a pre-defined width and to save them in separated data files. Each object can then be presented to a 2D classifier.

3. THE "GEOMETRIC" 2D CLASSIFIER

We have decided to develop an unsupervised method based on the Kohonen Maps (Kohonen 1997) without any parameterization of the objects.

3.1 The TAROT data

We saw in section 2 that our data processing software was able to extract all the objects of size not larger than a specified size. We choose a size of 11×11 pixels because most of our objects fit into this sub-image width. This difference is mainly due to the actual (or simulated) optical response of the instruments. The input vector is then a vector of 11×11 components. We speak indifferently about sub-image or input vector since each image is presented to the classifier as a vector of dimension 121. The sources can be of numerous types: stars, blended objects, galaxies, saturated objects.

3.2 The topological network

We use two dimensional arrays of 10×10 neurons. Each neuron *j* is associated with a weight vector W_j such as the component W_{ij} is connecting neuron *j* to pixel *i*.

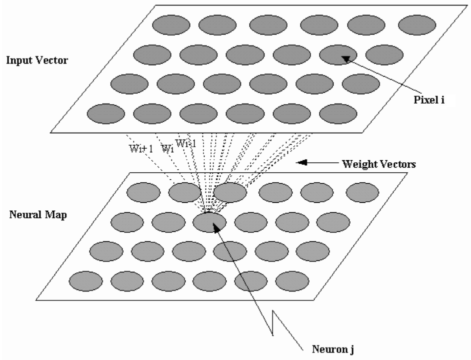

Figure 1: Configuration of the Kohonen Map

Before the training phase, initial values have to be given to the weight vectors. We have adopted a linear initialization (Kohonen 1997) where the weight vectors are initialized in an orderly fashion along the linear subspace spanned by the two principal eigenvectors of the input data used for the training phase. This initialization accelerates the training phase because the SOM is approximately organized in the beginning.

In each training step, for every new input vector I_i the network will compute the value D_j for each neuron:

$$D_j = \sum_{i=1}^{n} \left\| I_i - W_{ij} \right\| \qquad \textbf{(1)}$$

where n is the number of pixels, here 121, and $\left\| I_i - W_{ij} \right\|$ is the Euclidian distance. The Best Matching Unit (BMU) will be the neuron *c* for which the value of D_j is the smallest. This neuron is the one whose weights are the closest to the input data vector.

After having selected the winning neuron, we update the weight vectors of the neurons within a selected neighborhood according Equation 2:

$$W_{ij}(t+1) = W_{ij}(t) + h_{ci}(t) * (I_i - W_{ij}) \qquad \textbf{(2)}$$

where t denotes the time and h_{ci} is the neighborhood kernel associated to the winning neuron *c*. In order to preserve the topology of the data, neurons are connected to adjacent neurons by a neighborhood relation dictating the structure of the map. This relation is a decreasing function of time and of the distance of unit *j* to the BMU. It defines the region of influence that the input sample has on the map. We have tested the four available functions (bubble, gaussian, cut gaussian and epanechicov) and did not see any relevant changes in the results. For further works, we choose the Gaussian Kernel proposed by Kohonen (1997).

3.3 Application to Digitized Sky Survey images

Before testing our topological Network on real TAROT images, we wanted to test it on images used by Bazell and Peng (1998). These images were obtained by downloading Digitized Sky Survey (DSS) images in Flexible Image Transport System (FITS) Format from the Space Telescope Science Institute home page. We have downloaded the 60 galaxy and 27 star images used by Bazell and Peng (1998). Since we wanted to compare the results with the TAROT images, we retrieved images of size 11×11 pixels and with the same sampling as the TAROT images, i.e. 3”.88 pixel^{-1}.

With the topological neural network presented above, we then classify each of the 87 objects as a star or a galaxy by using the 86 other objects as the learning set. We have used the batch algorithm instead of the sequential because it is significantly faster with Matlab. Moreover, the training phase is performed in two phases, a first rough training phase (large initial learning rate and neighborhood radius) and a second fine tuning phase (small learning rate and neighborhood radius).

For this study, we use two kinds of standardization. The first preprocessing method (SOM1 in Table 2) consists in mean subtracted each input vector and normalized it to the unit length. The second method (SOM2 in Table 2) consists in dividing each component of the input vector by the square root of the sum of squares of all components.

We take Bazell and Peng notations and define the following quantities:

- The sensitivity is the total number of correctly identified galaxies divided by the total number of galaxies in the sample.
- The specificity is the total number of correctly identified stars that divided by the total number of stars.
- The Positive predictive number is the number of correctly identified galaxies divided by the number of identified galaxies by the network.
- The negative predictive is the number of correctly identified stars divided by the number of identified stars by the network.

Table 2 compares these quantities computed for our methods with the numbers obtained by Bazell and Peng for the Learning Vector Quantization (LVQ) and the Back-Propagation (BP) method .

	SOM1	SOM2	LVQ	BP
Sensitivity to galaxy	90%	88.3%	87%	97%
Specificity to galaxy	92.6%	96.3%	96%	96%
Positive predictive	96.4%	98.1%	98%	98%
Negative predictive	80.6%	78.8%	76%	93%

Table 2: Comparison of the results obtained by the method presented in this paper and the same quantities computed by Bazell and Peng (1998).

The results obtained with our topological neural classifier are comparable with the LVQ method results but are less encouraging than the Back-propagation one. In fact, if we do not consider objects that have their BMU on the frontier of the two clusters, our results can be improved.

If we compare the two normalizations methods, we can see that they provide similar results with a little preference for the second one, which is the fastest.

3.4 Application to TAROT images

We have applied this classifier on a set of images acquired by the TAROT observatory. A first work had provided interesting results by using a training set made of real TAROT images (Bringer & Boër 2000). This training set contained 5000 normalized data vectors representing background images and objects extracted by the processing software. Since the neighborhood function decreases as a function of the time, the images that will be presented first to the network will have a stronger influence on the map organization than the last ones. Hence we present the characteristic objects at the beginning of the training phase. The resulting U-Matrix is presented in Figure 2.

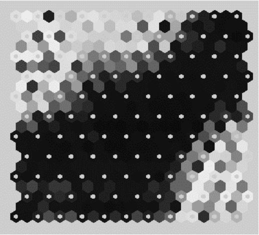

Figure 2: The U-matrix after training on real TAROT images

We clearly see three clusters that represent the different classes of objects that we had in the training set: background pixels, point source objects and extended sources. In the next step we present an object to the trained map and identify in which region is the corresponding BMU. If the wining neuron is in the frontier area, we can't give no more than a probability of being a point source.

These results are acceptable but we do not see any region corresponding to saturated objects or blended objects. In fact, there are not enough sources of this type in the training set to be adequately represented on the map. The training set does not seem to be adapted to a complete classification. Hence we have to look for a training set containing enough objects of each classes to be represented, including saturated points, extended sources, diffused absorbed regions and blended objects.

We looked for TAROT objects of this type and included them in the training set. The resulting U-Matrix is divided in six clusters corresponding to the different objects. This artificial addition of characteristics objects in the training set allows us to have a complete classification.

3.5 Discussion for the 2D classifier

The developed topological neural network is able to learn through experience and to discriminate between astronomical objects such as stars, galaxies, saturated objects or blended objects. The training set had to be enhanced to take into account objects that appear not frequently on our images. We paid a great attention to the type of sources presented first to the classifier during the training phase.

We used no parameterization because we did not want to loose object information. We have compared the results of our classifier with the one of Bertin and Arnouts (1996) on simulated images obtained by the Skymaker software. These images contain point sources and extended objects. We trained our classifier on these images and we analyzed the response to each object in a following test phase. The success rates for both classifiers are presented in Figure 3.

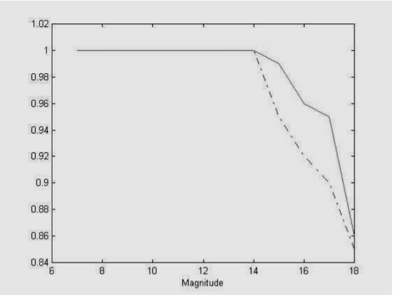

Figure 3: Classification success rate for our classifier (solid line) and the multilayer perceptron of Sextractor (dashed line – Bertin & Arnouts, 1996)

One can see that for faint objects, our classifier is better. In fact, the parameters used by Bertin and Arnouts (1996) are isophotal areas and do not allow a discrimination for faint intensity values. Since our method is able to discriminate correctly the different types of sources found in usual astronomical images, we tried to make a further step in analyzing temporal series of images.

4. DISCUSSION AND CONCLUSIONS

4.1 The temporal variability analysis

We are interested in the object luminosity evolution with the time. We have seen that a photometrical calibration with the USNO-A.2.0 catalogue was made during the processing phase so that we have the magnitude of the different objects relatively to a reference frame (differential photometry), as explained by Pojmanski (1997). The different time profiles obtained (light curves) may be of different types.

In Figure 4a and 4b, we present the light curves of two TAROT objects. All the 52 images analyzed to provide these light curves were taken with no filter (Clear position). The first object (Figure 4a) seems to be a variable star but we do not have enough data to know the type of variability. As a contrary, the second object (Figure 4b) does not present any variation and is then supposed to be a constant star at least to our precision level.

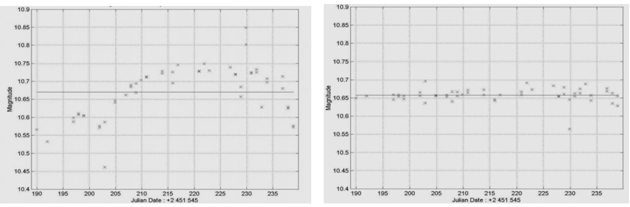

Figure 4a, 4b: Light curves of two objects detected on 52 TAROT images. 4a: TAROT J182157.9-143321.9 ; 4b: TAROT J203938.3-151324.6

To consider the objects temporal variability we can take directly the light curve and present it as a second signal to the topological neural network, the first signal being the sub-image of 11×11 pixels. It means that the sampling used for the curve plot should be the same for any object. However, in the case of astronomical imaging, it is often difficult to have large series of periodically sampled data, at least when the data is acquired over several nights: first the night/day alternance prevents from any regular sampling, second the weather conditions change from day to day, and third the visibility of the sources varies with their position in the sky and over the year. Hence we have to cope with the irregular sampling of astronomical data, and the impossibility to get the same sampling for all sources over the sky.

One approach is to use specific parameters of the light curve. We plan to analyze the curves in a frequency domain with the Fourier analysis and in the time-frequency domain with the wavelet analysis following the encouraging work presented in Szatmary et al. (1994). The results of this approach will be presented in a forthcoming paper.

4.2 Batch image analysis

The second way of investigation to take into account the temporal aspect is to use several sub-images centered on the same object but taken at different times.

A deeper image analysis is necessary to calibrate the different images. The method of difference image analysis (Alard & Lupton 1998) seems to be adapted to our problem. Through this method, we can obtain the light curves which would be a second way of constructing them and could confirm our first construction by photometrical calibration.

We plan to present the processed cubes (2D in space, 1D in time) of images to the classifier. The input vector would be of dimension 121 times the number of images composing the cube. However, the problem of the different cube (time) dimension will again appear because of the different data sampling of different sky regions.

We plan to work with temporal neural networks (Euliano & Principe 1996) to deal with this non-regular sampling problem.

4.3 Conclusions

We have introduced a Topological Feature Map able to discriminate between astronomical objects. The classification depends on the geometric characteristics of the sources. One of the advantages of the map is that we do not calculate any parameter and thus we do not introduce any bias on the objects we have to discriminate. A normalized sub-image centered on the source is presented to the network as a vector. A first approach enabled us to discriminate between stellar, non stellar and background objects. We have enhanced this work to deal with the classification of others sources such as saturated stars or blended objects.

The problem of the developed method is that we can only analyze objects that fit in a predefined size, here 11×11 pixels. Other objects, e.g. the larger ones, can't be classified. We have to find a solution and get over this problem to take into account all the detected objects.

After the elaboration of this 2D classifier, we are now looking forward to improve the map in order to deal with other types of astronomical objects. We are currently exploring the way of taking into account the temporal variability. The light curve analysis and/or the extracted parameters of this signal seem to be interesting inputs to use. Otherwise, we are trying to analyze cubes of images of the same source, taken at different times, as an entry of the 3D classifier. The main difficulty we have to deal with in this second step, is the irregular sampling of astronomical data.

The classification on the geometrical and temporal aspects is probably the next challenge of astronomical classification because it allows a more complete discrimination between astronomical sources.

REFERENCES

Alard C. & Lupton R.H. (1998). *ApJ*, 503, 325-331.

Bazell D. & Peng Y. (1998). *ApJ Supp. Ser.,* 47-55

Bertin E. & Arnouts S.(1996). *A&A Supp. Ser.,117, 393-404.*

Boër M. et al. (1999), *A&A Supp. Ser.*, 138, 579-580.

Boër M. et al. (2001), *A&A*, 378, 76-81.

Bringer M. et al.(2001). *Experimental Astrophysics*, accepted.

Bringer M. & Boër M. (2000). ADASS X, ASP Conference Proceedings, 216, 640.

Euliano N.R. & Principe J.C. (1996). *In Proceedings of IEEE ICNN*, 4, 1900-1905.

Kohonen.T (1997). Self-Organizing Maps, 2nd edition, *Springer Series in Information Sciences.*

Lutz R.K (1979). *The computer journal*, 23, 262.

Pojmanski G. (1997). *Acta Astronomica*, 47, 467-481.

Szatmary K., et al. (1994). *A&A Supp. Ser.*, 108, 377-394.