ASTRONOMY & ASTROPHYSICS
SUPPLEMENT SERIES

Astron. Astrophys. Suppl. Ser. **146**, 19–29 (2000)

An image database

III. Automatic extraction for millions of galaxies

G. Paturel, Y. Fang, C. Petit, R. Garnier, and J. Rousseau

CRAL-Observatoire de Lyon, F-69561 Saint-Genis Laval Cedex, France

Received February 17; accepted June 16, 2000

Abstract. This paper presents a method for extracting a catalogue of galaxy candidates from the Digitized Sky Survey (DSS). The method is based on a functional analysis applied on each individual plate. The standard deviation of pixel optical densities versus the inverse of surface area leads to a diagram in which extended and star-like objects are well separated. This diagram is used for a preliminary recognition. Then, a filtering process is applied using a Neural Network method associated with a training sample built with well identified objects. The main catalogue gives coordinates, total magnitude, isophotal diameter, axis ratio, position angle for 2 772 061 galaxy candidates. The method favors the detection of normal galaxies. This creates a bias against compact high surface brightness galaxies.

Key words: galaxies — catalogues — data analysis

1. Introduction

For several years we have been embarked on automatic galaxy extraction from various sources of images. Indeed, we are entering a new era, where the visual analysis will be replaced by an automatic one. MacGillivray et al. (1987) initiated automatic galaxy recognition with the COSMOS machine. A few years later similar techniques were used by Maddox et al. (1990) for the construction of the APM catalogue. Independently, Lauberts and Valentijn applied automatic surface photometry on galaxies discovered from a visual inspection. Nevertheless, we are still at the beginning of this process and new tools have to be invented. Some of the techniques used here were not even imaginable a few years ago.

Send offprint requests to: G. Paturel

Our general purpose is twofold:

- We aim to build large and complete samples of galaxies needed for extragalactic studies including the preparation of future radio or spectroscopic observations. For each galaxy we want to extract accurate position, diameter and axis ratio, position angle, magnitude and, if possible, some morphological description;
- We aim also to develop and test new methods of source recognition. This second target is important for forthcoming large surveys which will require automatic analyses.

In a preliminary study (Paturel et al. 1996) our target was limited to identification of galaxies already known in the LEDA[1] database and to extraction of some astrophysical parameters using our own digitization of the Palomar Sky Survey. The resolution of the digitization was too poor (6″) to allow recognition of new galaxies. We developed source extraction and automatic cross-identification algorithms.

In a second study (Vauglin et al. 1998) the same source extraction algorithms were used but we applied an automatic galaxy recognition based on Discriminant Analysis method. The I-band CCD images were obtained with the 1-meter ESO telescope for the Deep Near Infrared Survey (DENIS). Because of the high dynamic of the CCD receiver the separation between stars and galaxies is relatively easy. Stars have a very high central intensity and a small surface area while galaxies do not.

In the present study we are aiming at the most difficult task of recognizing new galaxies from the digitization of photographic plates (POSS1 and UK Schmidt plates). Because of plate properties the center of a source (star or galaxy) is generally saturated and only a few pieces of information can be derived from optical density of pixels[2]. Besides, the material is made of very inhomogeneous plates taken from very different regions. It is simply

[1] http://leda.univ-lyon1.fr

[2] Note that all along this paper the optical densities are defined according to the DSS documentation as

not possible to imagine that the same method will work in all conditions. This major conclusion of our preliminary analysis will force us to imagine a method adapted to each individual plate.

The different methods of automatic recognition can presently be classified in 5 classes as follows:

- The Discriminant Analysis (see Vauglin et al. 1998). It is one of the oldest automatic methods. Using a training sample, the sample is shared in classes according to a linear procedure which maximizes the inertia between classes and minimizes the inertia within a class;
- The functional analysis is similar to the previous method. The objects are plotted in the parameter space. The function sharing the different classes of the training sample is determined by an expert. It is not necessarily linear. In the present study we will use the functional analysis method as a first step of star/galaxy discrimination;
- The decision tree method (see Weir et al. 1995). Using a training sample, the discrimination power of some parameters is determined. Then, a combination of tests is carried out allowing the separation of the object sample into several classes;
- The neural network method (see: Storrie-Lombardi et al. 1992; Odewahn et al. 1992; Lahav 1994; Bertin & Arnouts 1996a,b). The network is a set of inputs (parameters) connected to outputs (classes of objects) through weighted, non linear links (neurones). Generally, using a training sample the weights are calculated in an iterative process to obtain the right output for a given set of input parameters. The neurones may be organized in several layers in order to increase the number of links. In the method developed by Bertin & Arnouts the training sample is automatically built from a proper model of stars and galaxies. In the present study, we will use the neural network method in a second step of discrimination;
- Kohonen charts (see Kohonen 1989) This method seems very promising because it does not require a training sample. The objects are classified using a kind of neural network for which the output classes are automatically defined. This method is similar to the cluster analysis. When the classes are built, an expert has to identify each of them with a given class of objects (galaxy, star, defect...). The classification of images was employed by Heydon-Dumbleton et al. (1989) for star galaxy classification for the Edinburgh-Durham Southern Galaxy Catalogue.

$d_c = 6553.4 \log(S_o/S)$, where S_o is the intensity of light transmission through an unexposed part of the plate and S is the transmitted intensity through the considered exposed part. We frequently use the term of pixel "intensity" for d_c instead of the proper term of pixel optical "density" which could lead to a confusion with the density in the sense of number of pixels per area unit.

The main characteristic of these methods is that they require a proper choice of parameters describing each object. This choice is not obvious. It is guided by the results obtained on a training sample for which each object has been classified by an expert. This is the main difficulty in the present application. The material (see Sect. 2) is so inhomogeneous that it would be necessary to build a training sample for each plate. This would mean classifying about 500 000 objects by eye. So, we worked in another way. In Sect. 3 we show that the diagram of the dispersion of pixel optical densities (i.e. standard error of the pixel intensities) versus the inverse of the surface area of a given object performs this Star/Galaxy separation well. Thus, we plotted these diagrams for each plate and made the separation between stars and galaxies by adopting a frontier function in an interactive manner. This constitutes the first step leading us to a preliminary catalogue of 4.3 million galaxy candidates and 47.4 million star candidates. In Sect. 4, we built a large, general training sample of 258 983 stars and 87 725 galaxies by cross-identifying our preliminary catalog with well established star and galaxy catalogs. This training sample was used to setup a neural network allowing us to filter the star/galaxy candidates. After this filtering step we got an all-sky catalogue of 3.2 million galaxy candidates. Finally, in Sect. 5 we made the internal cross-identification (what we call the *Auto-crossidentification*) for galaxies seen several times on different plates. Then, we made the cross-identification with LEDA galaxies and cleaned the catalogue in order to remove contamination by known extended objects (Planetary Nebulae, Globular clusters, Open clusters, Bright Nebulae, Bright Stars) and by very faint galactic stars. This led to the final catalogue of 2 772 061 galaxy candidates.

2. DSS material

2.1. Description of DSS scans

The material comes from the set of 102 compressed CD-ROM's of the Digitized Sky Survey produced at the Space Telescope Science Institute.

Each CD-ROM contains the digitization files of several plates. The Northern hemisphere ($\delta \geq +3$ deg) contains 644 plates, noted xe001 to xe643 and one additional (xe1001). The Southern hemisphere contains 894 plates noted s001 to s894. Each plate is constituted of 784 (28×28) individual scans of 500×500 pixels[3]. Each individual scan is labelled with a 2-symbol extension (from 0 to r), e.g., s828.00 to s828.rr. The scan *.11 is the most South-Eastern one, while the scan *.rr is the North-West one. We skipped all scans labelled with 1 or r (i.e., *.1x, *.x1, *.rx, *.xr) in order to reject the extreme edge of each plate. Thus, we processed about one million elementary scans. The size of a pixel is 1.70″. Each elementary

[3] Except for the last row which is 500×499.

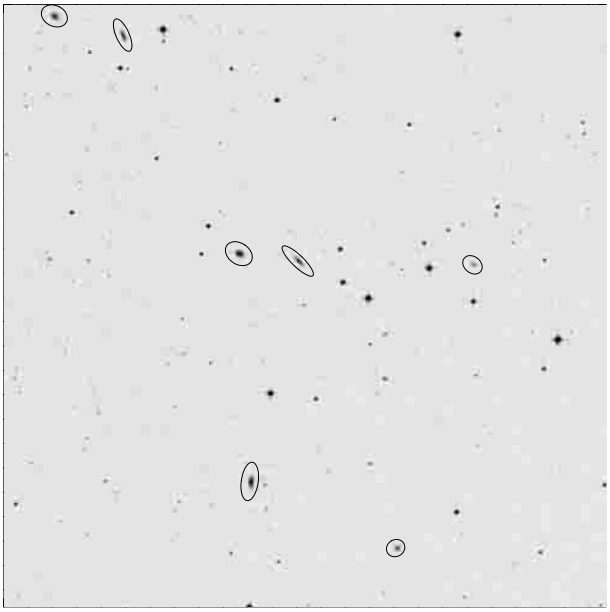

Fig. 1. A typical image (s828.ap) from the DSS CD-ROM's. This image is only 12.6 × 12.6 mm on the original plates. The frame is $14.2' \times 14.2'$. North is on the bottom side, East is on the left side. The seven galaxies clearly visible on the scan are new ones which were recognized by the automatic program

scan is about $14.2' \times 14.2'$. The Northern part is digitized from E(red) plates, while the Southern part comes from IIIa-J plates. An example of a typical elementary scan is given in Fig. 1.

2.2. Source extraction

The source extraction is made from uncompressed scans in the same way as in our previous papers (Paturel et al. 1996; Vauglin et al. 1999). The sky background is assumed to be homogeneous over an individual scan. Its mean intensity $I_{\rm bg}$ is calculated from the maximum of the histogram of pixel intensities. The standard deviation $\sigma_{\rm bg}$ is calculated by symmetrizing the low intensity side of the histogram. The threshold for source extraction is chosen as $I_{\rm bg}+3\sigma_{\rm bg}$. Only the sources having more than 36 pixels are kept. This means that the smallest objects have a size of $\sqrt{37}$ pixels, i.e., $10''$. Further we impose that the number of pixels on one side of the matrix of pixels is larger than or equal to three. If we note *npx*, the number of pixels per line and *nli* the number of lines, this means that $npx \geq 3$, $nli \geq 3$ and $npx.nli > 36$.

2.3. Equatorial coordinates

For each object we calculate the $\overline{x} - \overline{y}$ mean position of the matrix. The mean is obtained by weighting each pixel with its intensity. The J2000 equatorial coordinates are then calculated using the plate solution calculated at the Space Telescope Science Institute. The coefficients of the 13-th order polynomial solution are read in the header file associated which each plate. The internal accuracy is about $3''$ in right ascension and declination, near the equatorial plane. Near the Northern pole the accuracy is about $4''$. Near the Southern pole it is about $5''$. This is in agreement with the results found by a recent study of coordinate accuracy (Paturel et al. 1999; Paturel & Petit 1999). Most of the uncertainty comes from the positioning of the galaxy center.

2.4. Cleaning of sources

A treatment is then applied on matrices corresponding to overlapping objects in order to separate them into their different components. The basic assumption in this treatment is that astronomical objects have a central symmetry.

First of all, we determined the number of maxima in the pixel matrix. This is done as follows: when a maximum is found, the region around its position is inhibited and the following maxima are looked for, outside this region. The inhibited region around each maximum is actually the ellipse centered on the considered maximum and tangent to the nearest edges of the matrix.

The decomposition of a matrix in several matrices is then made in the following way: all pixels of the original matrix are considered one after the other. For a given pixel $P(i,j)$ we are searching for its symmetrical counterparts with respect to the n maxima M_k $(k = 1, n)$. If the symmetric counterpart of a given pixel $P(i,j)$, calculated with respect to the k-th maximum, is outside the matrix, or if it has an intensity below the sky background, the given pixel $P(i,j)$ is not attributed to the object re-constructed around this k-th maximum. If the pixel $P(i,j)$ belongs to several re-constructed objects, the pixel intensity is simply shared with equal weight between each object. No attempt has been made to share this intensity in a more refined way because the pixel intensities are not additive. The objects which result from this decomposition always have a central symmetry in accordance with our basic assumption. In the final catalogue a flag will remind us a given object results from such a decomposition process.

The matrices which are truncated by the edge of the scan are also extrapolated by symmetry if the maximum itself is not on the edge.

3. Preliminary analysis

We constructed 10 training samples in different regions and for different plates. For these samples the objects are classified, by eye, as *Star*, *Galaxy* and *Unknown*. From the corresponding matrix of pixels of classified objects we calculated many parameters and systematically plotted them

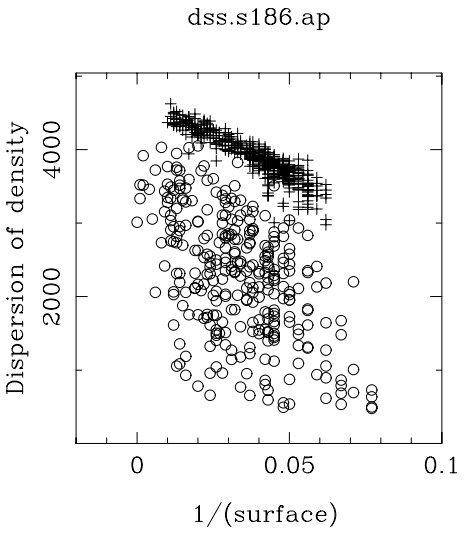

Fig. 2. Dispersion of density (i.e., standard deviation of the pixel intensities) of a given object versus the inverse of its surface area. Here the diagram is shown for a training field for which stars, galaxies and defects have been classified by human expert. Stars (crosses) and galaxies (open circles) are well separated. The units are the following: the dispersion of density $\sigma(d_c)$ is expressed in units of 6553.4 times the actual optical density ($\log S_o/S$) of the plate (see footnote in first page). The surface area is simply the number of pixels above the sky background (each pixel has a constant surface area of $1.7'' \times 1.7''$). The choice of these units is not crucial provided it is the same throughout the work as it is in the present paper

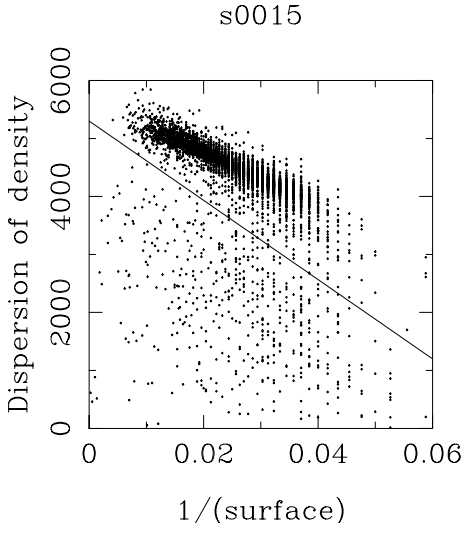

Fig. 3. Example of the diagram of dispersion of density versus the inverse of the surface area ($\sigma - 1/S$) for a field in the Southern equatorial hemisphere. The separation between stars and galaxies can be inferred from a comparison with Fig. 2. The separation curve between stars and galaxies is linear. The units are the same as in Fig. 2

two by two. We found that the dispersion of pixel optical densities (i.e. standard deviation of the pixel intensities) plotted versus the inverse of the surface area gives a diagram in which galaxies and stars are well separated in two distinct zones as in Fig. 2. The surface area is simply the number of pixels having an intensity I larger than the sky background intensity I_{bg}. The dispersion of pixel density, σ, is calculated as the standard deviation of the pixel intensities through the classical equation:

$$\sigma = \sqrt{\frac{n\sum I^2 - (\sum I)^2}{n^2}} \quad (1)$$

where the sums are calculated with the n pixels brighter than I_{bg}. An example of this diagram is given in Fig. 2.

3.1. First star/galaxy recognition

These diagrams were plotted for each plate (i.e., 1443 diagrams) and a polynomial separation curve was fitted manually to each of them. Three examples of these diagrams are given in Figs. 3 to 5, from the best to the worst. In Fig. 3 the frontier between Stars and Galaxies is a straight line. Stars and Galaxies are well separated. The frontier separating stars from galaxies is often quite linear as in Fig. 3, but not necessarily. Indeed, it is also common that the separation curve bends down for large objects (small $1/S$) as in Fig. 4. This seems to be due to the saturation of pixel intensities in either the central part of galaxies or in the halo of bright stars. This phenomenon has also been seen in the source extraction from I-band CCD images of the DENIS survey (Mamon, private communication). In regions of low galactic latitude the separation is more difficult as shown in Fig. 5 for a fields located at $b = -2$ deg. Thus, at low galactic latitude (i.e., $|b| < 18$ deg) the separation between stars and galaxies becomes more difficult especially for small objects (i.e. large values of the inverse of the surface area) because the two zones are progressively mixing with each other.

This first discrimination step produces a catalogue with 4 349 140 galaxy candidates and 47 352 280 star candidates. Hereafter, only the galaxy candidates will be considered. Nevertheless, our process did not remove every star from the galaxy candidate sample. A visual inspection showed that bright stars are sometimes counted as galaxy candidates because of their extended halo as explained above.

The construction of a completeness curve is a general way to check if a catalogue obeys the expected increase of object number with distance. If we assume that the number of galaxies within a sphere centered on the observer and of radius r increases as r^3, it can be shown that the number N of galaxies with an apparent diameter larger than a given limit D_{lim} follows the law: $\log N(D > D_{lim}) = -3 \log D_{lim} + cst$[4]. Generally, this completeness curve is used to check if a sample is complete

[4] The completeness curve expressed in apparent magnitude m can be written similarly as: $\log N(m < m_{lim}) = 0.6 m_{lim} + cst$. We will use this form in Sect. 5.4 when apparent

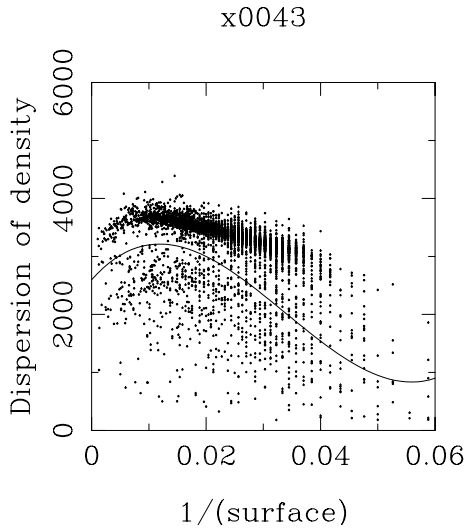

Fig. 4. Another example of a diagram of dispersion of density versus the inverse of the surface area ($\sigma - 1/S$) for a field in the Northern equatorial hemisphere. The separation curve between stars and galaxies is not linear. The units are the same as in Fig. 2

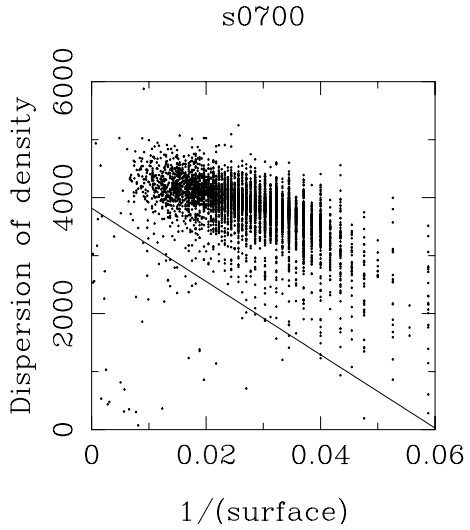

Fig. 5. Another example of diagram $\sigma - 1/S$ for a field located near the galactic plane ($b = -2\,\mathrm{deg}$) in the Southern equatorial hemisphere. The separation between stars and galaxies is more difficult. The units are the same as in Fig. 2

up to a given apparent diameter. Here, it is used to check if the number of galaxy candidates is homogenously distributed in space, as expected. Note that this curve is insensitive to the angular coverage of the catalogue or to the galactic extinction.

The over-sampling of large objects is confirmed by the completeness curve $\log N - \log D_*$ (Fig. 6), which shows an excess of large galaxies. Here, $D_* = \sqrt{4\,S''/\pi}$ is the equivalent diameter defined from the surface area S'' in

magnitudes will be calibrated. It was used by Hubble (1934). A demonstration is given, e.g., by Zwicky (1957).

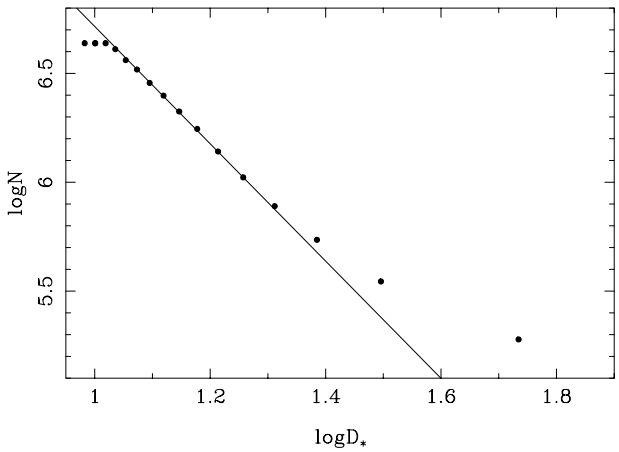

Fig. 6. Completeness curve $\log N - \log D_*$ built from the galaxy candidate catalog. The effective diameter D_* is expressed in arcseconds. The relation with the surface area S expressed in number of pixels is given by Rel. (2). It is visible that there is an excess of large objects (see text)

arcsec^{-2}. In this paper the surface area S is expressed in number of pixels. Thus, from the pixel size $1.7''$ it results:

$$\log D_* = 0.5 \log S + 0.283 \qquad (2)$$

where D_* is in arcseconds and the surface area S in number of pixels. We note that the completeness curve in diameter is quite linear for objects smaller than $\log D_* = 1.25$ (i.e. $1/S < 0.012$ or $D_* \approx 18''$). The completeness is fulfilled down to $\log D_* = 1.04$, i.e. $D_* = 10''$ in agreement with our cut-off (36 pixels). The catalogue contains one million galaxy candidates larger than $18''$, and thus 3.3 millions with diameter between $18''$ and $10''$.

Now, we have to clean up our galaxy candidate catalogue. This is the target of the next section.

4. Cleaning with a neural network

4.1. Construction of a large training sample

We use a Neural Network (hereafter NN) method to perform the cleaning. This requires the construction of a large training sample. We build it by cross-identifying each object of our preliminary catalogue with known stars or galaxies. The known stars are taken from the SAO catalogue. The known galaxies are taken from the LEDA database.

The cross-identification is based on the J2000 equatorial coordinates. The identity of two objects is accepted when there is only one object within a radius of $10''$. This severe constraint removes interacting objects which are not suitable for a training sample. So, we obtain 54186 objects classified as galaxy "G" and 90339 classified as star "S". Further, 2105 objects are classified as defect "D" because of their discrepant characteristics (e.g., a very elongated matrix with $nli/npx > 25$ or $npx/nli > 25$).

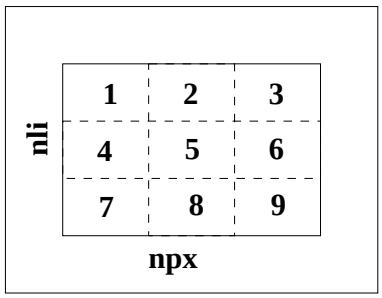

Fig. 7. Decomposition of a matrix in nine rectangles. This decomposition is used to define the diffraction cross parameter *dc* and the defect parameter *df*

Objects with $\log D < 1.25$ are not used for the construction of the training sample.

4.2. Definition of neural network input parameters

The NN has three outputs: G, S and D for galaxies, stars and defects, respectively. The choice of the input parameters of the NN is important. They must be very discriminant with reference to the outputs. There is no rule for the choice of these parameters. We define seven parameters which will be tested with our training sample:

1. P_1: The dispersion of pixel optical densities. This is the parameter defined for the first analysis (Sect. 2);
2. P_2: The inverse of the surface area. This is also the parameter previously used;
3. P_3: The logarithm of axis ratio of the object. An elongated object, if not detected as a defect, has more chance being a galaxy than a star; it is better to use the axis ratio of the object (calculated as in Sect. 5) instead of the ratio of the sides of the matrix (npx/nli or nli/npx). Indeed, a very elongated object can be located along a diagonal of the squared matrix;
4. P_4: The square of the external perimeter divided by the matrix surface area. This parameter is simply defined as $4(npx+nli)^2/(npx\,nli)$ and it is very sensitive to elongated features like scratches, or branches of a diffraction cross, or satellite tracks;
5. P_5: The ratio of the object surface area divided by the matrix surface area. This parameter is useful for detecting artefacts like those encountered on calibrating spots near the edge of the plate (in this case the ratio is nearly one), patchy objects or elongated features;
6. P_6: The diffraction-cross parameter *dc*. The matrix is divided into nine identical rectangles numbered from 1 to 9 according to Fig. 7. The diffraction cross is defined as the mean intensity of rectangles 2, 4, 6, 8 divided by the mean intensity of rectangles 1, 3, 7, 9;
7. P_7: The defect parameter. Let *na* (and *nb*) be the number of pixels with an intensity above (and below) the sky background intensity inside the central rectangle (5). The defect parameter is defined as: $df = (nb - na)/(nb + na)$. This is simply the proportion of blank pixels inside the central rectangle. For an astronomical object we expect no (or a few) blank pixel in the very center. For an extreme defect we have $df = 1$. For a normal object we have $df = -1$.

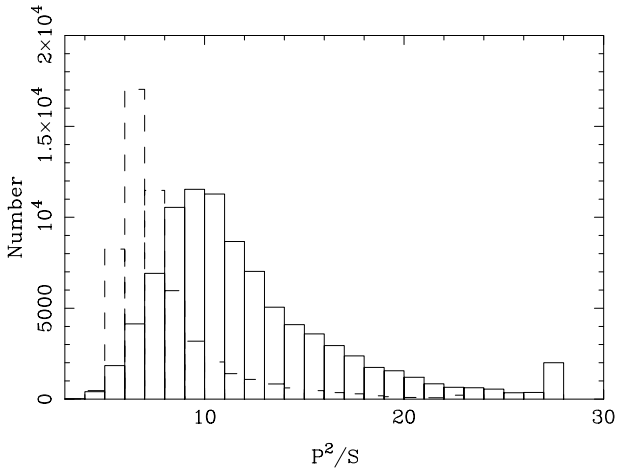

Fig. 8. Histogram of the square of the external perimeter divided by the matrix surface area for stars (solid line) and galaxies (dashed line)

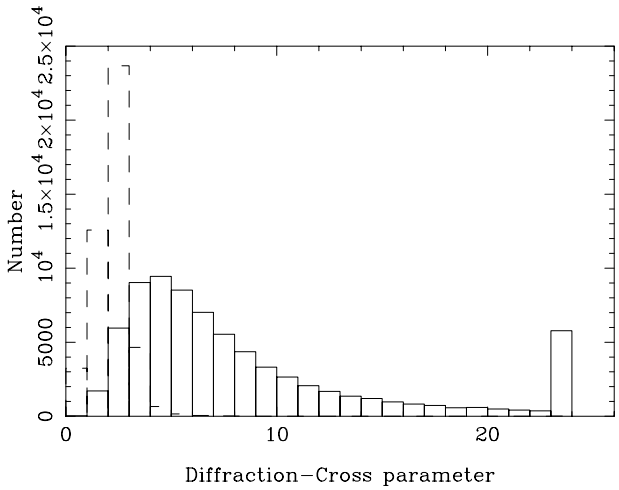

Fig. 9. Histogram of the diffraction-cross parameter for stars (solid line) and galaxies (dashed line)

In Figs. 8 to 10 we show some of the discriminant parameters for galaxies and stars.

4.3. The neural network definition

After several trials and errors, we adopted the NN represented in Fig. 11. Because there are only three output parameters (G, S, D) we adopted a simple NN with only one intermediate layer of 10 neurones each of them having 7 input parameters and 3 output ones. There are 100 free weights W connecting two neurones of two different layers. The input is a vector with seven components. The output

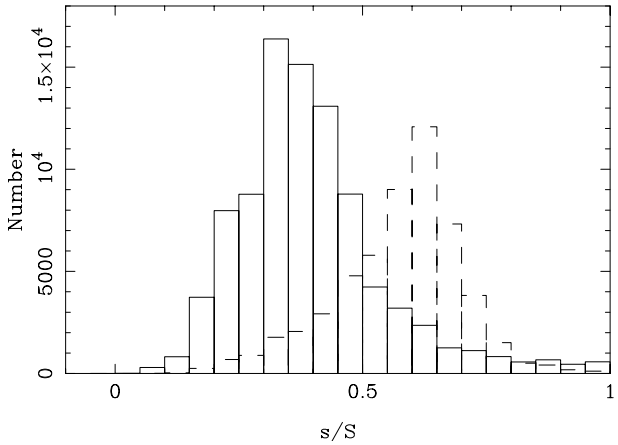

Fig. 10. Histogram of the ratio of the object surface area divided by the matrix surface area for stars (solid line) and galaxies (dashed line)

is a vector with three components. The expected output vectors are (1, 0, 0), (0, 1, 0), (0, 0, 1).

The different steps of the NN training are the following. First of all, the weights are randomly chosen between −1 and 1. Then, the training sample is read and each individual parameter is normalized by subtracting its mean and dividing by its standard deviation, both calculated from the whole sample. So, we obtain seven input components P_i. An object is a vector with seven components. Then, for each object, the seven input parameters P_i are entered and propagated down to the last layer (output layer). For this purpose, the input X of a given neurone is the weighted mean of its input connections while its output is calculated through a non linear sigmoid function[5]:

$$s = \frac{1}{(1+\mathrm{e}^{-X})}. \quad (3)$$

The error vector E is determined by comparing the calculated output vector with the known output vector (this is done with the training sample). Then, E is propagated back using the weights for sharing the error onto the different branches and the derivative of the sigmoid function

$$\frac{\mathrm{d}s}{\mathrm{d}X} = \frac{-\mathrm{e}^{-X}}{(1+\mathrm{e}^{-X})^2} \quad (4)$$

for crossing back a neurone. Finally, the weights are corrected accordingly.

The process is repeated (i.e., the normalized input parameters are entered and the calculation is done again) until the system becomes stable. In practice this is done by testing different iteration numbers.

We did some preliminary tests on the whole training sample to find the best number of intermediate neurones and the best number of iterations. We tested the number

[5] The choice of a sigmoid function is justified by the fact that we are looking for a bimodal answer. Note that this sigmoid function is not applied on inputs P_i which are not considered as neurones.

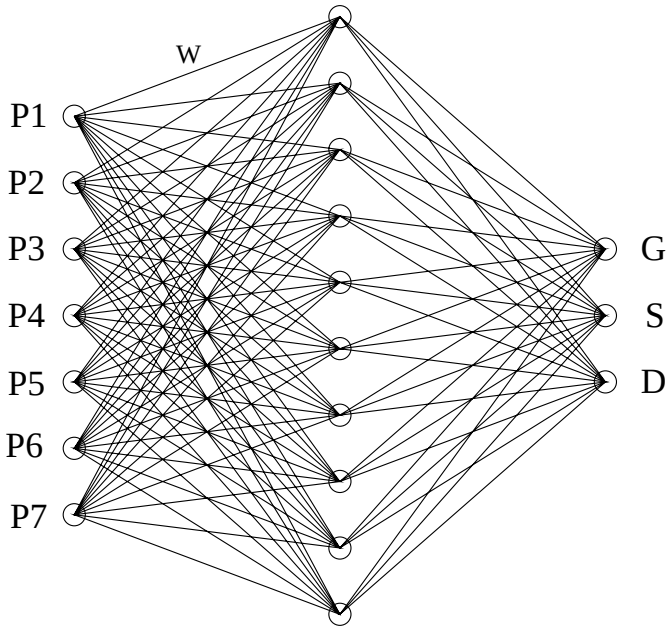

Fig. 11. Representation of the adopted neural network. $P1$ to $P7$ are the seven input parameters. G, S, D are the three output values for galaxy, star and defect, respectively. Each open circle is a neurone. The connection between two neurones has a weight W

of intermediate neurones between 7 and 42 and the number of iterations between 50 and 600. Finally, we decided to adopt 10 intermediate neurones and 100 iterations.

4.4. Setting and testing the NN

An efficient way to demonstrate the success of an automatic classification programme is the usage of a "control sample", i.e. determine the automated parameters (G, S, D from NN) in the same way for the control sample and compare these with independent reference values of the control sample. Actually, we built nine control samples. The whole sample of 132972 objects with proper object classification was divided into ten non-overlapping subsamples $S0$ to $S9$ having the same size (1/30-th of the total sample). The NN was programmed ten times and we kept the solution ($S0$) giving the best result for the whole sample. Then, to prove the validity of the NN, configured with $S0$ only, we applied this configuration to the nine independent samples $S1$ to $S9$. The results for these nine control samples are given in Table 1.

Obviously, the components of the calculated output vector (G, S, D) are not exactly 0 or 1. The component G obtained with the training sample is shown in Fig. 12. Most of its values are 0 or 1 (i.e., the NN answers either "yes" if it is a galaxy, or "no" if it is not a galaxy).

In our control we considered a result as good when the largest component corresponds to the expected one. For instance: if we got the answer: $G = 0.7$, $S = 0.6$, $D = 0.1$

Table 1. Application of the NN, configured with a subsample $S0$, to nine independent subsamples $S1$ to $S9$, for which the result of the classification is known. When the NN gives the good answer the result is considered as a success. We give the size and the percentage of successes for each subsamples $S1$ to $S9$ and for the whole sample

samples	size	percentage of success
$S1$	4667	94%
$S2$	4667	94%
$S3$	4667	93%
$S4$	4667	93%
$S5$	4667	93%
$S6$	4667	92%
$S7$	4667	94%
$S8$	4667	94%
$S9$	4667	94%
Total sample	132972	84%

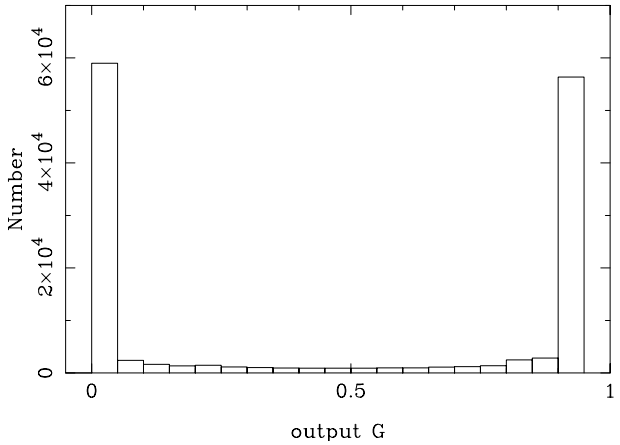

Fig. 12. NN-Output G obtained with the training sample. Most of the values are close to zero or one. Components S or D have exactly the same bimodal distribution

for an object known as a galaxy ($G = 1$, $S = 0$, $D = 0$) we concluded that the NN gave a right answer, because the largest component is G.

For the final application of the NN we imposed more severe constraints in order to reduce contamination of the galaxy catalogue by stars or defects. The adopted conditions were those given in Table 2. Further, some objects are considered a priori as defects when the parameter P_5 (ratio of the object surface area by the matrix surface area) is larger than 0.95 (case of a matrix almost without sky background pixel), or when the axis ratio is larger than 100.

Using this NN cleaning we classified: 1 147 332 objects as galaxies (G), 134 509 as probable galaxies (g), 1 940 573 as "possible galaxies" (-). We classified: 179 842 objects as stars (S) (in addition to the catalogue of 47 million stars previously extracted), 946 884 as defects (D).

Table 2. Additional constraints

Conditions	Classification	code
$G \geq 0.9$ and $S < 0.5$ and $D < 0.5$	Galaxy	G
$G \geq 0.8$ and $S < 0.2$ and $D < 0.2$	Probable Gal.	g
$E \geq 0.8$ and $D < 0.5$ and $G < 0.5$	Star	S
$D > E$ and $D > G$	Defect	D
otherwise	Possible Gal.	-

5. Construction of the main catalogue

At this stage we have a catalogue of 3 222 414 galaxy candidates. We now have to make the "auto-crossidentification" to merge a same object seen on different plates. Because the information on the original plate will be lost in such a merging process we have to apply now the corrections which are plate-dependent, like the effects of the mean airmass extinction or the distance to the center of the chart (Rousseau et al. 1996; Garnier et al. 1996).

5.1. Astrophysical parameters

We use a Principal Component Analysis method applied on pixels positions (i, j) of the matrix associated to an object. So, we derive for each object a 2×2 covariance matrix from which we calculate its eigenvalues (v_1 and v_2) and corresponding eigenvectors. The position angle $\beta_{\rm DSS}$ of the major axis is determined from the direction of the first eigenvector (eigenvector associated with the highest eigenvalue). The major and minor axes are deduced from the square root of the first and second eigenvalues. The apparent magnitude is deduced from the sum of all pixel intensities. We thus obtain the following parameters:

- Position angle $\beta_{\rm DSS}$ measured from North towards East. Note that the position angle is calculated first with respect to the edge of the frame and then corrected for the actual direction of North.
- Major axis diameter
$$\log D_{\rm DSS} = \log \sqrt{v_1} + C_1 \quad (5)$$
where C_1 is a constant.
- Axis ratio
$$\log R_{\rm DSS} = \log \sqrt{v_1/v_2} + C_2 \quad (6)$$
where C_2 is a constant.
- Apparent magnitude
$$m_{\rm DSS} = -2.5 \log \sum_{i,j} (I(i,j) - I_{\rm bg}) + C_3 \quad (7)$$
where C_3 is a constant. Magnitudes are corrected for the mean atmospheric extinction (assuming that the plate is taken at the meridian) and for the distance to the center of the plate according to Garnier et al. (1996):
$$\Delta m_{\rm DSS} = -0.0006 \Delta r - c \sec \zeta. \quad (8)$$

Where Δr is the distance of the considered galaxy to the plate center (in mm), ζ is the zenithal distance and c the atmospheric extinction coefficient (0.2 or 0.1 depending on the plate).

In order to calibrate these equations we extracted from the LEDA database the apparent blue diameter D_{25}, the major to minor axis ratio $R_{25} = D_{25}/d_{25}$ (axes are defined at the isophote 25 mag arcsec^{-2}) and the total $B_{\rm T}$ magnitude for the galaxies of the training sample. These quantities are in the system of the Third Reference Catalogue (RC3, de Vaucouleurs et al. 1991). We get the following results (σ is the standard deviation, n is the number of remaining objects after 3-σ rejection:

$$\log D_{25} = \log D_{\rm DSS} + 0.09 \quad \sigma = 0.06 \quad n = 5619 \tag{9}$$

$$\log R_{25} = \log R_{\rm DSS} + 0.00 \quad \sigma = 0.07 \quad n = 5555 \tag{10}$$

$$B_{\rm T} = m_{\rm DSS} + 30.6 \quad \sigma = 0.34 \quad n = 5544. \tag{11}$$

It is worth noting, that the accuracy is reasonably good owing to the rather rough comparison. The standard errors should thus be considered as upper limits for galaxies up to $\approx$ 15-th magnitude. These magnitudes will be reanalyzed in a future work in order to take into account local and secondary effects.

For stars, a comparison with SAO magnitudes gives a preliminary calibration:

$$m(\text{stars}) = m_{\rm DSS} + 25.1 \quad \sigma = 0.46. \tag{12}$$

The difference of zero-points for stars and galaxies suggests that the dispersion of pixel optical densities intervenes in a refined calibration.

5.2. Auto-crossidentification

The catalogue of 3 222 414 galaxies is sorted according to the declination (the search is easier and faster with such a sorting). Each galaxy is compared with all the others. This is done four times because one galaxy may only appear four times at the intersection of four charts. At each of these four iterations only the locally two closest galaxies are merged if their separation calculated along a great circle is smaller than a given limit. This procedure avoids the result depending on the order the galaxies are considered (in other words, the merging is done according to a physical measurement but not following an arbitrary order). The limit of the separation is calculated from the actual uncertainty on the position:

$$d_{\rm lim} = d_{\rm o}(1 + \cos^2\delta)^{1/2} \tag{13}$$

where $d_{\rm o}$ is the nominal uncertainty on coordinate measurement along one direction. We adopt $d_{\rm o} = 6''$ (Paturel et al. 1999). In practice, the galaxies are not actually merged at this stage. They simply receive an internal numbering. A galaxy which appears several times receives the same internal number. The merging will be done later. Two reasons justify the postponement of the merging: 1) For each occurrence of a given object we have a matrix. It is not easy, without loss of information, to merge these matrixes in one mean matrix. 2) The crossidentification with LEDA will be done for each extracted object, even if it appears several times. This will give us a chance to detect possible inconsistency (e.g. a galaxy identified once with a given LEDA galaxy and then with another one when it is extracted from another plate).

Thus, we will still work with the catalogue of 3 222 414 galaxy candidates (a direct merging would have lead to a catalogue of 2 876 111 galaxies. No inconsistency is found).

5.3. Cross-identification with LEDA

In view of this cross-identification we carried out a campaign of measurement of accurate coordinates. More than 34000 positions of LEDA galaxies were measured (Paturel et al. 1999; Paturel et al. 2000) and we studied the accuracy of the coordinates provided to us by large catalogues (Paturel & Petit 1999). We added some recent accurate measurements (Cotton et al. 1999). After this work we have a list of 194544 galaxies from LEDA with accurate coordinates and the main astrophysical parameters (diameter, axis ratio, position angle and magnitude).

The cross-identification is based essentially on coordinates using a method similar to the one used for the auto-crossidentification. Nevertheless, two modifications are introduced: 1) The limit of the separation is calculated from the previous formula (Rel. 13) but the value of $d_{\rm o}$ is deduced from the weighted mean of the coordinate accuracy (Paturel & Petit 1999) and quadratically increased by the uncertainty of the DSS coordinates ($6''$), because the coordinates we are comparing have independent errors (this was not the case for auto-crossidentification). 2) When several galaxies match the position criterion we use astrophysical parameters to choose the best one. For this purpose we calculate a generalized separation between the objects according to

$$t = \frac{1}{N}\sum_{i=1}^{N} w_i \frac{|\Delta X_i|}{\sigma(X_i)} \tag{14}$$

where w_i is the weight assigned to each parameter X_i (e.g., coordinates, diameter, axis ratio, position angle). ΔX_i is the difference of the parameter X_i for the two galaxies in test. After some trials we assigned a weight of 7 for coordinates, 1 for diameter, 2 for axis ratio $2\log R$ for the position angle. After this step 144 721 objects are identified in LEDA (corresponding to 107 991 galaxies because some objects appear several times). Thus, 86 553 galaxies known only in LEDA are added, leading to a catalogue of 3 308 967 galaxies before the merging of repeated galaxies (or 2 962 664 galaxies if the merging is done). The added objects are galaxies fainter than the magnitude limit of the DSS, or low surface brightness galaxies not detectable by our program, or a few very large objects (larger than individual frame).

At this stage we build the mean catalogue where a galaxy appearing several times is merged into one object. There is no practical difficulty because each object has its internal number from the auto-crossidentification step. Nevertheless, we must take into account that some periodical parameters (like the right ascension or the position angle) must be treated with special care. For instance, two measurements of the position angle of a galaxy elongated in the N−S direction may produce, e.g., 175 deg and 5 deg. The mean of both measurements is not 90 deg but 180 deg. After having merged all objects appearing on different plates we obtain a catalogue of 2 876 111 objects.

5.4. Removal of non extragalactic extended objects

Automatic program of galaxy recognition cannot differentiate a true galaxy from, e.g. a planetary nebula or a globular cluster. Further, filaments in a bright nebula, in a HII region, in the neighborhood of a very large galaxy or in the halo of a very bright star can well be recognized as a galaxy. In order to remove such artefacts we constituted a catalogue by collecting objects prone to create them. This catalogue of "forbidden zones" is built from the following objects:

- Bright stars (code ST) brighter than $m_v = 7$ mag taken from the SAO catalogue (17405 objects);
- Galaxies (code GA) larger than $5'$ taken from LEDA (311 objects);
- Globular clusters (code GC) taken from Harris & Racine (1979) (160 objects);
- Open clusters (code OC) taken from Lynga (1983) (1151 objects);
- Bright Nebulae (code BN) taken from Lynds (1965) (1125 objects);
- HII regions (code H2) taken from Sharpless (1959) (626 objects);
- Planetary Nebulae (code PN) taken from Acker (1992) (1143 objects).

For stars the forbidden zone is the central circle (diameter $D_s = 1.1'$) and the branches of the diffraction cross. The total extension (with both arms) of one branch is estimated to $B = -3m_v + 25$ (arcmin). For galaxies, the forbidden zone is the surface of the ellipse defined by its axes D_{25} and d_{25} (at the isophote 25 mag arcsec^{-2}) and the position angle of the major axis β (from North towards East). For all other objects the forbidden zone is the surface of the object assumed to be circular of diameter D. The forbidden zone catalogue gives for each object: the code (ST, GA, GC, OC, BN, H2, PN), the right ascension and declination for equinox 2000, and the parameters for the definition of the forbidden zone (D_s and B for stars, D_{25}, d_{25} and β for galaxies and D for others). This catalogue is sorted according to declination and contains 21921 objects.

Table 3. Number of rejected galaxy candidates

Code	Object	Number of rejection
ST	Stars ($m_v > 7$)	4028
GA	Galaxies ($> 5'$)	34 017
H2	HII regions	25 560
GC	Globular clusters	1906
OC	Open clusters	12 578
BN	Bright Nebulae	112 318
PN	Planetary Nebulae	196
	Total	190 603

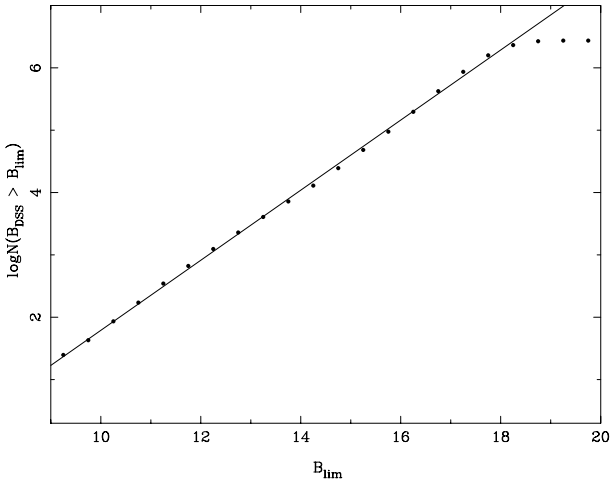

Fig. 13. Completeness curve for the main catalogue of 2 772 061 galaxies. The completeness is fulfilled up to 18.2 mag

In Table 3 we give the number of rejected objects for the different classes of forbidden zones. After these cleaning and merging steps 2 772 061 galaxies remain. This catalogue constitutes the main catalogue from which we will start to work. A completeness curve made with these 2 772 061 galaxies shows that the completeness limit is about 18.2 mag (see Fig. 13).

The slope of the linear part is 0.56 ± 0.01. This is significantly less than the theoretical value (0.6). This result has been permanently found and has been interpreted in several ways (fractality, incompleteness, flat distribution of galaxies).

6. Conclusion

After many visual inspections it appears that the method is very efficient for detecting faint galaxies and for determining their proper size and orientation: low surface brightness galaxies are well detected; near the galactic plane many galaxies seen by the eye have been automatically identified and many new galaxies have been discovered. This constitutes an incredible improvement even if some star-like galaxies are missed and if some artefacts remain.

The catalogue has been loaded into the LEDA database. It considerably changes the management of our database because in a given field all galaxies brighter than 18^{th} mag are present at their right place. A code has been assigned to each galaxy to measure the reliability of the detection. Progressively, the database will be cleaned for remaining artefacts. Presently, 1 million galaxies can be accessed through LEDA and the corresponding catalogue is available electronically.

We will use the catalogue itself with the matrices of galaxies for several astrophysical purposes:

- Calculation of accurate inclinations from disk flateness (for spiral galaxies);
- Automatic and impersonal morphological classification;
- Estimation of local variation of the galactic extinction from galaxy counts;
- Definition of environment of a galaxy;
- Analysis of structures (clusters and groups);
- Analysis of correlation functions.

As a by-product we got a catalogue of 50 millions stars. It is to be noted that among these stars there are compact extragalactic objects. Some of them will be recovered from special analysis in preparation.

Acknowledgements. This work was based on photographic data obtained using The UK Schmidt Telescope. The UK Schmidt Telescope was operated by the Royal Observatory Edinburgh, with funding from the UK Science and Engineering Research Council, until 1988 June, and thereafter by the Anglo-Australian Observatory. Original plate material is copyright (c) the Royal Observatory Edinburgh and the Anglo-Australian Observatory. The plates were processed into the present compressed digital form with their permission. The Digitized Sky Survey was produced at the Space Telescope Science Institute under US Government grant NAG W-2166. The amount of calculation needed by the present study was so large that we were obliged of borrowing many computers during hollidays. We thank Ph. Prugniel, F. Simien, E. Pecontal, E. Emsellem, H. Di Nella, L. Copin for lending their computers. We also thank the Pole Scientifique de Modélisation Numerique (PSMN) for giving us a privilagiate access to their computing facilities. We thank Pierre Valvin, Kristine Bonnefoy and Christelle Lamy-Charrier for helping us to write a neural network program. We also thank Isabelle Frechet for contributing to visual tests of the method and Mikko Hanski for contributing in the large scale analysis.

References

Acker A., 1992, The Strasbourg-ESO Catalogue of Galactic Planetary Nebulae. Garching bei München, European Southern Observatory

Bertin E., Arnouts S., 1996a, A&A 311, 356

Bertin E., Arnouts S., 1996b, A&AS 117, 405

Cotton W.D., Condon J.J., Arbizzani E., 1999, ApJS (to be published)

Garnier R., Paturel G., Petit C., Marthinet M.C., Rousseau J., 1996, A&AS 117, 467

Guide Star Catalog, 1989-1992, Space Telescope Science Institute (GSC)

Harris W.E., Racine R., 1979, ARA&A 17, 241

Heydon-Dumbleton N.H., Collins C.A., MacGillivray H.T., 1989, MNRAS 238, 379

Hubble E.P., 1934, ApJ 79, 8

Kohonen T., 1989, Self-Organization and Associative Memory. Springer-Verlag, Berlin, ISBN 0-387-51387-6

Lahav O., 1994, in Vistas in Astronomy, special issue on ANNs in Astron. 38, 3

Lauberts A., Valentijn E.A., 1989, The Surface Photometry Catalog of the ESO-Uppsala Galaxies, ESO, Munich

Lynga G., 1983, Revised Catalogue of Open Cluster Data, Lund Observatory

Lynds B.T., 1965, ApJS 12, 163

MacGillivray H.T., Dodd R.J., Beard S.M., 1987, Proceedings No. 28, "Astronomy From Large Databases", Murtagh F., Heck A. (eds.). Garching

Maddox S.J., Sutherland W.J., Efstathiou G., Loveday J., 1990, MNRAS 243, 692

Odewahn S.C., Stockwell E.B., Pennington R.L., Humphreys R.M., Zumach W.A., 1992, AJ 103, 318

Paturel G., Petit C., 1999, A&A 352, 431

Paturel G., Garnier R., Petit C., Martinet M.C., 1996, A&A 311, 12

Paturel G., Petit C., Garnier R., Prugniel Ph., 1999, A&AS 140, 89

Paturel G., Petit C., Garnier R., Prugniel Ph., 2000, A&AS 144, 475

Rousseau J., Di Nella H., Paturel G., Petit C., 1996, MNRAS 282, 144

Sharpless S., 1959, ApJS 4, 257

Storrie-Lombardi M.C., Lahav O., Sodré Jr. L., Storrie-Lombardi L.J., 1992, MNRAS 259, 8

de Vaucouleurs G., de Vaucouleurs A., Corwin H.G. Jr., 1976, Second Reference Catalog of Bright Galaxies. University of Texas Press, Austin (RC2)

de Vaucouleurs G., de Vaucouleurs A., Corwin H.G. Jr., Buta R.J., Paturel G., Fouqué P., 1991, Third Reference Catalog of Bright Galaxies. Springer-Verlag (RC3)

Vauglin I., Paturel G., Borsenberger J., Fouqué P., Epchtein N., Kimeswenger S., Tiphène D., Lanoix P., Courtois H., 1999, A&AS 135, 133

Weir N., Fayyad U.M., Djorgowski S., 1995, AJ 109, 6

Zwicky F., 1957, in "Morphological Astronomy". Springer-Verlag, Berlin