AUTOMATIC MORPHOLOGICAL CLASSIFICATION OF GALAXIES

ZSOLT FREI
Institute of Physics, Eötvös University, Budapest, Pázmány P. s. 1/A, H-1117, Hungary;
E-mail: frei@alcyone.elte.hu

Abstract. Sky-survey projects under development or in the early phase of data collection will soon provide large numbers of high-quality CCD images of galaxies. We have developed a software-system for automatic morphological classification of the many galaxies in such data sets. Our steps include: automatic removal of foreground stars, deprojection of galaxies, converting images to polar coordinates (to count arms through FFT) and reduction of many signatures representing the galaxy on the image. We have developed a galaxy-catalog of 113 nearby galaxies to test our software. Some of the above parameters, together with a numerical type index and absolute magnitude found in the literature for our galaxies are subjected to principal component analysis. We find some good correlations between pairs of data, confirm findings of previous studies, and propose further statistical analysis of the data presented here.

1. Introduction

We are currently engaged in a long term project of developing an automated mechanism for morphologically classification of galaxies. In this work we present the first step towards fully automatic classification. This step is 'first' in a sense that the techniques presented here should be further tested on large amounts of well understood data before they can be reliably applied to large surveys. The techniques described here are *fully automatic*. All data reduction and image processing steps, and subsequent analysis of signatures extracted from the images has been integrated together in a software system which can provide a morphological type as an output from a two-dimensional input image (or images, if data were taken in more than one photometric bands).

Automatic classification of galaxies – besides being an interesting challenge for a long time – is crucial for large digital sky surveys. There are several survey projects currently planned, and some of them are already in the late stages of construction and software design. The amount of data they will provide is much larger than astronomers have had to cope with before, a fact which is due largely to the nature of these surveys: they are all computer driven automatic systems designed for continuous data collection, processing and archival. Perhaps the largest set of data will come from the Sloan Digital Sky Survey (SDSS), which will use a dedicated 2.5 meter special-purpose telescope to scan the Northern Galactic cap for 5 years (Gunn and Knapp, 1993). An imaging survey will produce a detailed map

Astrophysics and Space Science is the original source of publication of this article. It is recommended that this article is cited as: *Astrophysics and Space Science* **269–270:** 577–583, 1999.

of the sky in five photometric bands up to mag. 23. Approximately 10^8 galaxies will be cataloged, and based on this map about 10^6 galaxies will be selected for a spectroscopic survey with fiber-fed spectrographs on the same telescope.

The primary purpose of our efforts presented in this paper is to provide automatic classification techniques which will be useful for the SDSS project, a step in its atlas-image pipeline. Our aim is to develop algorithms which are robust and fast enough to perform classification on the SDSS spectroscopic survey galaxies, and perhaps be useful in a more limited way up to the photometric limit.

2. The Signatures

We have developed a procedure to analyze images step-by-step. Quantitative parameters are obtained through these steps, along with measures of error, and indicators of confidence levels. Given a two dimensional image of a galaxy, we first deproject it to face-on orientation, and then explore the spiral structure in the disk (if the galaxy, indeed, has a disk). Obviously, for distant galaxies, this is difficult to do, but during the first few steps at least a good indication of a bulge-to-disk ratio and a concentration index can be obtained.

The most valuable information for classification is in the radial brightness profile. For large galaxies the profile along the major or minor axes can be used with moderate success, but for smaller images the information in all pixels should be utilized. Therefore deprojection is necessary not only to study the spiral structure, but we found it to be the most useful and robust way of obtaining the radial brightness profile. This profile is obtained through averaging pixel values in concentric annuli, and then an analytical model composed of an exponential disk and a de Vaucouleurs bulge is fitted.

The face-on disk is transformed to polar coordinates, the direction and the strength of the spiral arms are computed, and a one dimensional (1D) profile across the spiral arms is constructed. The Fourier components of this profile give us information about the modes of the spiral arms. Measures of the small scale power in the image is obtained from the representation in polar coordinates. Almost all of these steps are repeated for another image of the same galaxy taken in another photometric band. The weighted mean of the scale lengths of the bulge and disk components are calculated, and the amplitude of these components are fitted for both images. Measurements of the total light corresponding to the components are used to compute the color of the bulge and the disk separately.

2.1. Position and inclination angles

The determination of the position angle using intensity-weighted second-order image moments seems to be the most robust technique here. In order to minimize the possibility of the moments being affected by strong spiral arms we compute the

moments on the autocorrelation (AC) image corresponding to the original image. The AC image will have a strong signal in the direction of the position angle of the galaxy on the source image, but will be less affected by spiral arms and other smaller scale sources (HII regions, etc).

We assume that inclination angles (for lenticular, spiral and irregular galaxies) can be calculated from the ratio of the major and minor axes of the image, and this ratio will be the actual measure of *inclination* we use. The image-moments can be used to calculate the axis ratio as well.

We developed another technique, which may be even less affected by small scale features in the image. This method is called 'silhouettes', since we compute the silhouettes of the image of the galaxy as it would appear looking at it along the major and minor axes, and determine the inclination by matching the scale length of the two silhouettes in the least squares sense.

Our estimates of the position and inclination angles are certainly good enough to construct a radial brightness profile from the deprojected galaxy, since the radial profile and the parameters we obtain from it (concentration index, disk-to-bulge ratio) are not too sensitive to the ellipticity of the image.

2.2. BULGE-DISK DECOMPOSITION AND CONCENTRATION INDEX

The next step is to construct a radial brightness profile using all pixels in the image. We do this by averaging all pixels in concentric annuli (centered on the photometric center of the galaxy). Once the profile is constructed we fit a two-component model to match this profile. The brightness of the bulge component follows the well known de Vaucouleurs $r^{1/4}$ law. The disk is simply an exponential disk. We used the multidimensional downhill simplex method for fitting and restarted the fitting routine three times to be sure of convergence.

A disk-to-bulge ratio is obtained from the fitted parameters by dividing the total light coming from the two components. The concentration index is also obtained from the brightness profile.

Our concentration index is from Kent (1985): $c \equiv 5\log(r_{80}/r_{20})$, where 20% and 80% of the total light is emitted within r_{20} and r_{80}, respectively. Theoretically, $c = 2.80$ for an exponential disk-profile, and $c = 5.25$ for a de Vaucouleurs bulge. The uncertain determination of the 'total light' and the limited radial resolution in the brightness profile are certainly sources of error, but we obtained values of c very close to these predicted by theory for artificial galaxies composed of only an exponential disk or an $r^{1/4}$ bulge.

2.3. $\theta - \ln r$ PICTURES AND SPIRAL ARMS

The spiral structure of deprojected disk galaxies can be easily explored in polar coordinates. We construct an image with the polar angle θ horizontal and with the natural logarithm of the radius, $\ln r$, on the vertical axis and the measured intensity

in the original image (or some processed version thereof) at those coordinates at the appropriate pixel in the image.

If the spiral structure of a given disk-galaxy is logarithmic we expect to see straight arms in the $\theta - \ln r$ picture. We can determine the tangent of these lines formed by the arms, thus obtaining the opening angle of the spiral arms. Spiral arms can be obtained more robustly from the AC image of the $\theta - \ln r$ picture.

The correct determination of the lower and upper bounds of the region of the $\theta - \ln r$ picture to be used for calculating the AC image is very important. If the lower bound is low, tips of a central bar, or an incorrectly deprojected or subtracted bulge may all affect the AC image. If the upper bound is too high, above the disk cutoff, noise will affect the AC image greatly because the radial normalization of the $\theta - \ln r$ picture results in very high magnification of the noise in the outer regions, where the signal is negligible. Determining these bounds based on the fitted scale length of the disk (see previous section) is the most successful approach.

Using the direction of the spiral arms obtained from the AC image, we skew the $\theta - \ln r$ picture, so that the arms are vertical. The image can be collapsed vertically to get a 1D profile of the arms. The 1D Fourier transform of this profile is calculated. There is usually very little power left in modes higher than 8, all due to noise, not real arm structure.

The Fourier transform is a complex function, and the absolute value *only* does not contain all the information. Moreover, it is difficult to 'read off' the relative importance of the arms corresponding to different modes, since if, say, a spiral structure contains a narrow, prominent two-arm component, the upper harmonic modes (4, 6, 8, etc) will show a considerable amount of power. We developed an *ad-hoc* method to modify the power distribution to include information from the phases, and depict better what an observer actually sees looking at the 1D profile, or the picture of the galaxy. Not all of these steps may be physically meaningful, but this is a well determined, repeatable procedure, and if it is useful for learning something about the given galaxy, this use justifies the procedure.

The arm structure, especially if weak or distorted may not give enough information about the galaxy. We wish to obtain independent measures of the fine scale structure in the disk. Fine scale power comes not only from spiral arms, but from all other light concentrations, such as HII regions, bright stars, etc.

One measure is the interquartile range of the pixels in the $\theta - \ln r$ picture, in the region from which the AC image is calculated. We obtain the lower and upper quartiles of the pixel values, and divide the difference by the median. A similar, and perhaps less robust measure of the power on small scales is the standard deviation of the pixel values, divided by the mean, over the same region. The standard deviation is divided by the mean of the data to obtain the 'relative' small scale power. The latter is much more likely to be affected by small structures such as un-removed foreground stars and HII regions, than the former.

Another – obvious – measure is derived from the Fourier transform of the 1D profile. The relative power of mode 0, compared to the total power (all modes

TABLE I

Correlation matrix for five important parameters (see text)

A. Correlation Matrix

vari.	DB	$c.i.$	M_{gal}	c_{gal}	T	mean	s.d.	$\sum(cc)^2$	angle
DB	1.00					0.474	0.633	1.20	0
$c.i.$	–0.81	1.00				3.174	0.619	1.40	177
M_{gal}	0.08	–0.11	1.00			–19.182	1.575	0.05	272
c_{gal}	–0.50	0.58	–0.09	1.00		0.746	0.162	1.13	175
T	0.53	–0.63	0.15	–0.73	1.00	2.732	3.461	1.23	350

B. Eigenvalues and Eigenvectors

p.c.	eigenv.	cumul.	DB	$c.i.$	M_{gal}	c_{gal}	T
ξ_1	2.915	(58%)	–0.489	0.521	–0.108	0.481	–0.497
ξ_2	0.982	(77%)	–0.114	0.077	0.989	0.048	0.026
ξ_3	0.663	(91%)	0.570	–0.394	0.081	0.551	–0.457
ξ_4	0.264	(96%)	–0.141	–0.092	–0.061	0.679	0.712
ξ_5	0.176	(100%)	0.635	0.748	0.008	0.035	0.191

summed) is an important measure of the strength of the arms. This measure, however, requires that the arm detection and the skewing of the $\theta - \ln r$ picture is successful, and is a reliable measure only if the spiral structure is coherent radially.

3. Principal Component Analysis Of Properties

We have developed a galaxy catalog to test the automatic procedure described above. We obtained many parameters for the 113 galaxies in our catalog, and subjected these parameters to principal component analysis (PCA) to find correlations among the parameters and to obtain few components which best describe the morphologies of galaxies.

PCA is useful for understanding correlations among linear combinations of parameters in multidimensional space, a task which is impossible to perform by calculating correlations simply between pairs of parameters. If p parameters are observed of n galaxies, the parameters can be plotted in a p dimensional Euclidean vector-space (the total of n vectors, with p components each). PCA is capable finding directions along which data varies the most in this space (the *first* space). In simple words, PCA is an orthogonal rotation of the original unit vectors in the euclidean vector-space, so that the new unit vectors point along the largest variances in the data. PCA is only capable of finding linear combinations, and

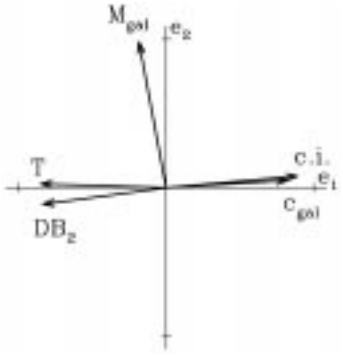

Figure 1. CV diagram of the five parameters from Table I.

consequently it is always suggested to transform all the variables so that they have similar dynamic ranges. In our case this is done by taking the logarithm of the disk-to-bulge ratio.

Previous studies suggested good correlation between the *Hubble type* (T) and parameters derived from the radial brightness profile: the *concentration index* ($c.i.$) and the *disk-to-bulge ratio* (DB) (see Okamura *et al.*, 1984; or Kent, 1985). It is also known that galaxies define a one-parameter sequence in color-color space, and *color* (c_{gal}) correlates with the Hubble type (Larson and Tinsley, 1978). Whitmore (1984) found no correlation between the Hubble type (the 'form' of the galaxies) and the *absolute magnitude* (M_{gal}) of galaxies. First we explored these five parameters (italicised above), and added the rest of the parameters we obtained from the images for a full study.

The correlation matrix for the five parameters is shown in the upper half of Table I. The mean and the standard deviation of each parameter is displayed. The sum of the squared correlation coefficients for each variable ($\sum(cc)^2$) is also shown. This is a useful measure of how much the given variable is correlated with the other variables in the table. The last column in the upper table is the angle between the first variable and the given variable projected to the plane of the first two principal components. The eigenvalues and the eigenvectors are shown in the lower half of Table I. Each row begins with the designation of the principal component (ξ), and it is followed by the eigenvalue, the cumulative variance accounted for by the given principal component and all of those above it, and the components of the normalized eigenvector.

To display the relationship graphically among the parameters we will use unrotated correlation vector (CV) diagrams (for a good description see Whitmore (1984)). It is apparent from the CV diagram that the other four parameters form a plane perpendicular to the direction of M_{gal}. The correlation coefficients between M_{gal} and the rest of the parameters are very low. It is also apparent that there is a good (anti-)correlation between T and c_{gal} (and between DB and $c.i.$). The correlation coefficients among the two sets of parameters (T and c_{gal} vs. DB and $c.i.$)

are around 0.5. This is an important result of this study, since it tells us how much correlation there is between the (subjective) Hubble type and the concentration index, or disk-to-bulge ratio. It was earlier suggested that these parameters correlate well with the type T, and indeed, 0.5 is not a negligible correlation coefficient.

Adding the rest of the parameters we obtained from the images we concluded: (a) the three largest correlation coefficients between T and another parameter are those with c_{gal}, $c.i.$ and DB in decreasing order; (b) the anti-correlation with the length of the arm is also significant (–0.58); and (c) the rest of the coefficients are below 0.35. We found good correlations between the concentration index and the disk-to-bulge ratio, and also two parameters representing the small scale power in the disk. We showed that these two sets of parameters do not correlate well, and are therefore independent descriptions of the galaxy morphology.

We decided to explore the correlation between T and DB and $c.i.$ further. We found that a linear combination of these two parameters corresponds best with numerical Hubble type.

4. Conclusions

We have presented new methods of analyzing photometric images of nearby galaxies. We described methods to decompose disk galaxies, to construct the radial brightness profile, and to compute the bulge-to-disk ratio and the concentration index from it. We have developed techniques to obtain many other parameters describing detail in the disk.

In the previous section we performed a principal component analysis of some selected parameters of the catalog galaxies. We also demonstrated the known good correlation between the numerical Hubble type and the color of the galaxy, and confirmed previous results that the absolute magnitude does not correlate well with the rest of the parameters representing the galaxy. We constructed a new parameter out of the disk-to-bulge ratio and the concentration index (reducing the variance) and show good correlation with the numerical Hubble type. We suggest that further studies of additional parameters and larger data-set could result in a few combination of parameters which are robust and objective descriptors of galaxy morphologies.

This research was supported in part by OTKA through grant no. F017150 and grant no. F029243.

References

Gunn, J.E. and Knapp, G.R.: 1993, in: B.T. Soifer (ed.), *Sky Surveys: Protostars to Protogalaxies*, *ASP Conference Series* **43**, p. 267.
Frei, Z., Guhathakurta, P., Gunn, J.E. and Tyson, J.A.: 1996, *Astron. J.* **111**, 174.
Kent, S.M.: 1985, *Astrophys. J. Suppl.* **59**, 115.
Okamura, S., Kodaira, K. and Watanabe, M.: 1984, *Astrophys. J.* **280**, 7.
Larson, R.B. and Tinsley, B.M.: 1978, *Astrophys. J.* **219**, 46.
Whitmore, B.C.: 1984, *Astrophys. J.* **278**, 61.