Parallel 2D local pattern spectra of invariant moments for galaxy classification

Ugo Moschini[1], Paul Teeninga[1], Scott C. Trager[2] and Michael H.F. Wilkinson[1⋆]

[1]Johann Bernoulli Institute, and [2]Kapteyn Astronomical Institute,
University of Groningen, P.O. Box 407, 9700 AK Groningen, The Netherlands
{u.moschini,m.h.f.wilkinson,s.c.trager}@rug.nl
{p.teeninga}@home.nl

Abstract. In this paper, we explore the possibility to use 2D pattern spectra as suitable feature vectors in galaxy classification tasks. The focus is on separating mergers from projected galaxies in a data set extracted from the Sloan Digital Sky Survey Data Release 7. Local pattern spectra are built in parallel and are based on an object segmentation obtained by filtering a max-tree structure that preserves faint structures. A set of pattern spectra using size and Hu's and Flusser's image invariant moments information is computed for every segmented galaxy. The C4.5 tree classifier with bagging gives the best classification result. Mergers and projected galaxies are classified with a precision of about 80%.

Keywords: classification, astronomy, pattern spectra, parallel computing

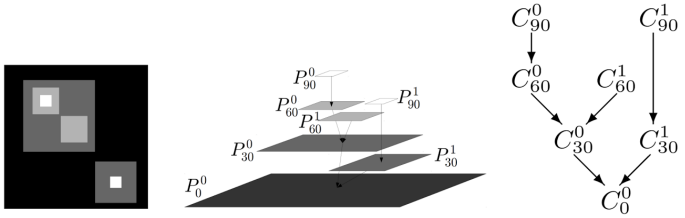
1 Introduction

Nowadays, astronomy, as well as many other scientific disciplines, has to face the problem of analysing the burden of data that are produced by modern instrumentation and tools. In particular, sky surveys like the Sloan Digital Sky Survey [1] (SDSS) contain hundreds of millions of objects. Finding and classifying the relevant objects, mostly stars or galaxies, cannot be done manually. The classification of the morphologies of galaxies is not a trivial task. Parametrized models [14], non-parametric approaches [11] and crowd-sourcing projects such as GalaxyZoo [10] are used for galaxy classification. Commonly used morphological classes are elliptical, spirals and *mergers*. The galaxies of the latter type are irregular and asymmetrical galaxies often connected by faint filaments of dust or gases, whose length and shape varies according to the stage of the merging. Parametrized models assume that the galaxies show a predefined light distribution and they are not irregular. Crowd-sourcing takes time and if a different classification class arise, users must be asked again to give their feedback. Non-parametric approaches are independent of any assumption and are often effectively combined with machine learning techniques [2, 11].

In our work, we want to distinguish mergers from other galaxies that could look close to each other due to a projection effect but are not interacting. They are referred

⋆ This work was funded by the Netherlands Organisation for Scientific Research (NWO) under project number 612.001.110.

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.



(a) 2D image (b) Peak components (c) Max-Tree

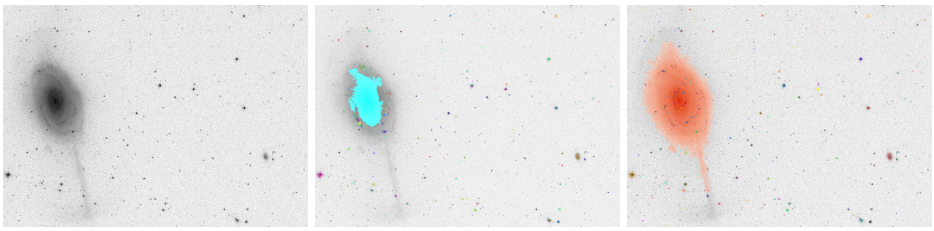
Fig. 1: A grey-scale 2D image with intensities from 0 to 90, its peak components P_h^k at intensity h and the corresponding max-tree nodes C_h^k.

to as *projected* or overlapping galaxies. We apply a non-parametric approach and investigate if a tool from mathematical morphology can be used for classifying galaxies: the pattern spectrum. Pattern spectra were introduced by Maragos in [12]. A pattern spectrum can be defined as an aggregated feature space that shows how much image content is present in the image components that satisfy certain classes of attributes. It represents the distribution of image details over those classes. Experiments with 2D pattern spectra using size and shape classes showed that they work effectively in pattern recognition and classification tasks on popular data sets [21]. Image invariant moments from Hu [9] and Flusser [7] have also been applied successfully to many pattern recognition tasks, ranging from satellite imagery to character recognition. The term *local* pattern spectra is used when pattern spectra are computed for every segmented object and not on the image as a whole. In this paper, we selected a dataset from SDSS Data Release 7 containing 196 merging and overlapping galaxies. The method in [20] is used to segment the galaxies and retain the faint tidal structures typical of mergers. Local pattern spectra of size and image moment invariants are created for each galaxy. Sets of 2D local pattern spectra are computed in parallel, modifying a parallel algorithm presented in [15]. Such collection of pattern spectra is used as feature vector in C4.5, a decision tree classifier. Section 2 reviews the segmentation method, Section 3 and Section 4 describe the moment invariants used and define the local pattern spectrum. In Section 5 and Section 6 show the experiments performed with different pattern spectra and the speed performance of the parallel algorithm. Mergers and projected galaxies in the dataset are correctly classified in about 80% of the cases.

2 Max-tree object segmentation

Any grey-scale image can be represented as a set of connected components, that are groups of pixels path-wise connected and with the same intensity, according to the classical definition of connectivity [18]. There being an ordering in the image intensities, the connected components can be nested in a hierarchical tree structure, namely a max-tree [17]. Every node in the tree corresponds to a peak component, which is a connected component at a given intensity level in the image. The leaves of the tree represent the local maxima of the image. Fig. 1b illustrates the hierarchy of peak components at

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.



(a) Original image (log scaled) (b) Segmentation by SExtractor (c) Max-Tree segmentation

Fig. 2: (a) original image with a galaxy whose protruding filament and outer boundary are segmented better in by our method [20] in (c) than by SExtractor in (b).

different intensities h for the image in Fig. 1a. The arrows in Fig 1c represent parent-child relationships that link the nested peak components. Useful measures related to the components can be computed efficiently while the max-tree is being built. The node structure is augmented with the attributes that can be used to identify the nodes that would possibly belong to objects of interest. In the specific case of astronomical images, in [20] we proposed a novel method that performs astronomical object detection using the max-tree structure. It starts with estimating the background in the image looking for tiles devoid of objects. After the background is subtracted, a max-tree is built. A statistical attribute filtering [5] is used. It is based on the expected noise distribution in the image compared with the distribution of the power attribute, as a function of its area. It selects which nodes of the tree are likely to belong to objects and which nodes are due to noise. The method showed an improved segmentation with respect to Source Extractor [3] (SExtractor), especially on faint extended sources, as in Fig. 2. The background estimate of SExtractor often correlates with astronomical objects. This is an issue in the case of structure close to the background level. If such structures are considered background, there is no threshold value able to identify them. On top of that, a fixed threshold above its background estimate is used to identify objects on a highly quantized version of the image, without considering noise and object properties. We refer to [20] for more examples and a detailed explanation of the differences between SExtractor and our solution.

3 Moments

Moment invariants are properties used to characterise images, for classification and pattern recognition tasks. In this paper, moment invariants are computed through geometric (raw) moments, for each connected component rather on the image as a whole. Let us define a component at a given intensity level as a binary image $f(x, y)$, where the background is made of the pixels that do not belong to the component. The moment of order $p + q$ of f is defined as:

$$m_{pq}(f) = \sum_{(x,y)\in f} x^p y^q f(x, y). \tag{1}$$

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

Raw moments can be transformed in *central* moments by using the coordinates of the component centroids $\bar{x} = m_{10}/m_{00}$ and $\bar{y} = m_{01}/m_{00}$ as follows:

$$\mu_{pq} = \sum_{(x,y)\in f} (x - \bar{x})^p (y - \bar{y})^q f(x, y). \quad (2)$$

Central moments are translation invariant. *Normalised central* moments are scale invariant moments derived from central moments, defined as $\eta_{pq} = \mu_{pq}/\mu_{00}^{\alpha}$, with $\alpha = p + q/2 + 1$. Hu [9] derived seven two dimensional descriptors suitable for 2D images (or components). The seven invariant moments were demonstrated to be translation, scale and rotation invariant. They are defined in terms of normalised central moments below:

$\phi_1 = \eta_{20} + \eta_{02}$
$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$
$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$
$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})$
$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}))(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)$
$\phi_6 = (\eta_{20} - \eta_{02})((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$
$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) + (3\eta_{12} - \eta_{30}(\eta_{21} + \eta_{03})(3(\eta_{30} + \eta_{12})^2) - (\eta_{21} + \eta_{03})^2)$

Flusser in [7] and Flusser and Suk in [6] showed that Hu's moment invariants are dependent and incomplete. It was pointed out that in pattern recognition problems, it is important to work with independent descriptors because they grant the same discriminative effect at the lowest computational cost, especially in high dimensional feature spaces. There are six Flusser's invariant moments that form a complete and independent set. They correspond to a subset of five Hu's moments ($\psi_1 = \phi_1$, $\psi_2 = \phi_4$, $\psi_3 = \phi_6$, $\psi_5 = \phi_5$, $\psi_6 = \phi_7$) with the addition of ψ_4:

$$\psi_4 = \eta_{11}((\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21}^2)) - (\eta_{20} - \eta_{02})(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21})$$

The six Flusser invariants used are of the second and third order. The first moment invariant ψ_1 is know also as the normalized moment of inertia and it can be used as a measure of the elongation of a component. We recall here that the moments ψ_4 and ψ_6 are skew invariant in the sense that they can separate between mirrored components, that it is not always desirable. As in [21], we modified the segmentation algorithm to compute raw moments for every node. The moment invariants can be easily computed from raw moments when the spectra are created. In such way, it is possible convey in the pattern spectrum of an object the information coming from the moment invariants for all the components that a galaxy is made of. Hu's and Flusser's moment invariants are used in the computation of the pattern spectrum of every galaxy, previously segmented with the algorithm illustrated in Section 2.

4 Parallel local pattern spectra

Without a hierarchical image representation like the max-tree, pattern spectra are computed using a number of morphological openings, with structuring elements of increasing sizes. The difference between two consecutive openings is an image that contains

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

Type of Moments	Invariants	Area of components	Dimension of PS
Hu	$\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7$	$>= 1$	7x7
Flusser	$\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6$	$>= 4$	14x14
Flusser*	$\psi_1, \psi_2, \psi_3, \psi_5$		30x30

Table 1: Pattern spectra with different settings are used. The first two columns illustrate three sets of moments considered; the third column shows that components with area larger than or equals to 1 and 4 pixels were used for different pattern spectra; the fourth column shows the number of bins used for area and moment invariant values.

the structures (connected components) having a size in the range given by the areas of the two structuring elements. The sum of the pixel intensities in the difference image, gives the amount of image detail for those components. In this case, the pattern spectrum is a 1D histogram where each bin corresponds to a range of areas, called *size* pattern spectrum. Max-trees can be used to compute the pattern spectra efficiently and independently of the structuring element used to sample the image. The nodes of the tree contain the exact area of every component without explicitly computing morphological openings. On top of that, they keep track of useful attributes such as moments, for all the components. Multidimensional pattern spectra can be obtained by binning the connected components not only according to their area but also to some other attribute. For example, a measure of the elongation of a component given by the first Hu's invariant moment was used in [21] was used to generate a 2D *shape-size* pattern spectra. A pattern spectrum is called *local* when it is computed on a segmented object and not on the whole image. After the max-tree is built, for each node visited in arbitrary order, the bin of the spectrum in which a given attribute value falls is chosen. Once the correct bins are identified, the product between the area of the component and the intensity difference with its parent node in the tree is added at that location. The computation of the pattern spectra can be parallelized applying a technique similar to the one presented in [15], tested on high resolution remote sensing images. When the pixels are partitioned among the threads of the parallel program, each thread handles the nodes of the tree falling under its partition corresponding to those pixels. The difference is that now a pattern spectrum for every object must be stored, whereas in [15] a single pattern spectrum was used for the whole image. Every thread stores now a number of pattern spectra equals to the number of objects found. The pattern spectra values are computed in every thread and the partial results are merged at the end. The output is a list of pattern spectra, one for every object in the image.

5 Classifying galaxies: the experiment

We investigate if 2D local pattern spectra that use invariant moments show statistically significant differences to distinguish merging from projected galaxies. The images used are obtained from the SDSS Data Release 7 [1]. The data set used consists of 98 monochrome $r-$band images containing a pair of close-by galaxies each for a total of

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

98 mergers and 98 overlapping galaxies. The image resolution is 2048x1489 pixels. Galaxies classified as interacting were selected from the Arp's Atlas of Peculiar Galaxies, whereas the projected galaxies were obtained from the classification of GalaxyZoo. The Weka 3.6.12 software package [8] and its implementation of the C4.5 decision tree classifier algorithm [16] were used to perform the experiments. The feature vector used as input by the C4.5 algorithm is a collection of 2D pattern spectra: one for each moment invariant. For every local 2D pattern spectrum, a dimension shows the bins for the area values of the components. The area is normalized dividing by the total size of the segmented galaxy. For every galaxy, the node of the tree that corresponds to its component of lowest intensity is normalised to have area equal 1. It represents the perimeter of the galaxy. Early tests showed that a logarithmic binning of area gives better results: it is desirable to have finer bins for lower values of area. The other dimension of the pattern spectrum refers to the bins for the moment invariant values. A 2D pattern spectrum is created for every moment invariant. For example, in the case of Hu's invariants, a set of seven 2D pattern spectra will be used as feature vector. The moment invariants tested are summarised in the first two columns of Table 1 and reported below.

Hu: The seven Hu's moment invariants.
Flusser: The six Flusser's moment invariants of the second and third order.
Flusser*: A non-skew invariant subset of the Flusser's invariants.

The value of such invariants was computed for all the components belonging to the 196 objects in the dataset. In total, there are about $2.5 \cdot 10^6$ components for all the 196 objects. As reported in the third column in Table 1, tests were also performed discarding the components smaller than 4 pixels: about $2.0 \cdot 10^6$ nodes were left. Binning of moment values was chosen so that the same number of components falls in each bin. Moment invariant and area values were binned in 7, 14 and 30 intervals, as reported in the last column of Table 1. The feature vector of every galaxy is made of a number 2D pattern spectra equal to the number of moment invariants used. For example, in the case of Hu's moment invariants, seven pattern spectra are calculated, each one with dimension, for example, 7x7 or 30x30. In this case, every galaxy is described by a 7x49 or 30x210 feature vector. As mentioned in Section 4, once the correct location in the pattern spectrum is identified given an area and a moment invariant value, the area of the component is multiplied by the intensity difference with its parent node: such product is added at that location in the spectrum. As a further test, a normalised version of the pattern spectra is computed: the product of area of a component and intensity difference is normalized by dividing for the total area of the galaxy. In short, a total of 36 different kinds of feature vectors were tested by composing three types of moment invariants, components with area larger than or equal to 1 and 4 pixels, three different binnings and lastly enabling or not normalization of the pattern spectra values.

6 Classification results and speed performance

The C4.5 classifier in Weka software package was tested in three variants: standard, with adaptive boosting (AdaBoost) and with bagging, commonly used techniques to improve decision tree classifiers. The results are shown from Table 2 to Table 5. The

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

Moments	Dim. of PS	C4.5 (%)	Boosting (%)	Bagging (%)
Hu	7x7	71.41	76.34	79.64
Hu	14x 14	71.88	73.65	78.11
Hu	30x30	64.71	67.80	72.80
Flusser	7x7	72.48	78.05	81.00
Flusser	14x14	72.00	72.51	76.54
Flusser	30x30	67.83	68.41	70.53
Flusser*	7x7	74.93	78.17	79.87
Flusser*	14x14	71.87	72.61	75.56
Flusser*	30x30	67.70	67.24	70.50

Table 2: Percentages of correctly classified instances. Pattern spectra were normalized and components with area $>=1$ were processed.

Moments	Dim. of PS	C4.5 (%)	Boosting (%)	Bagging (%)
Hu	7x7	74.52	78.33	78.88
Hu	14x 14	70.67	73.86	77.33
Hu	30x30	66.83	66.15	71.44
Flusser	7x7	71.36	77.21	79.60
Flusser	14x14	71.17	73.23	77.92
Flusser	30x30	66.91	67.72	72.04
Flusser*	7x7	73.62	77.26	79.54
Flusser*	14x14	72.51	72.41	77.46
Flusser*	30x30	66.58	65.72	71.82

Table 3: Percentages of correctly classified instances. Pattern spectra were normalized and components with area $>=4$ were processed.

last three columns of every table show the percentage of correctly classified instances (galaxies) for the three variants. Every value is the average result got over 10 repetitions of 10-fold cross validation, for a total of 100 folds. The tables refer to the four cases that originate from the analysis of components larger than or equal to 1 or 4 pixels and normalizing or not the pattern spectra values. The average standard deviation of all the runs is 9.92 for the standard C4.5 and 8.90 for the C4.5 with bagging. The best results are achieved with bagging enabled in Table 2, Table 3 and Table 5 and with boosting in Table 4. In every table, the three highest percent values are underlined. We notice that there is no big difference among the correct predictions over the tables. The two highest percentages of correct classification of objects as merging and overlapping galaxies are 81.00% in Table 2 for the Flusser set normalized, and 80.89% with the Flusser set not normalized in Table 4, using 7x7 pattern spectra. In general, better results are got when a smaller number of bins is chosen and when all the components are considered, not only those larger than 3 pixels. This could be explained by the fact that merging galaxies often show faint structures made of small dust-like particles, that result in an increased number of small components. Smaller components might convey a kind of information

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

Moments	Dim. of PS	C4.5 (%)	Boosting (%)	Bagging (%)
Hu	7x7	71.64	80.08	79.11
Hu	14x 14	71.88	73.65	78.11
Hu	30x30	70.61	70.06	76.63
Flusser	7x7	73.15	80.89	79.79
Flusser	14x14	72.00	72.51	76.54
Flusser	30x30	70.56	69.21	73.17
Flusser*	7x7	74.42	80.13	78.70
Flusser*	14x14	71.87	72.61	75.56
Flusser*	30x30	69.00	68.38	72.61

Table 4: Percentages of correctly classified instances. Pattern spectra were not normalized and components with area $>=1$ were processed.

Moments	Dim. of PS	C4.5 (%)	Boosting (%)	Bagging (%)
Hu	7x7	74.52	78.33	78.88
Hu	14x 14	70.67	73.86	77.33
Hu	30x30	65.67	68.96	74.41
Flusser	7x7	71.36	77.21	79.60
Flusser	14x14	71.17	73.23	77.92
Flusser	30x30	67.02	66.58	73.66
Flusser*	7x7	73.62	77.26	79.54
Flusser*	14x14	72.51	72.41	77.46
Flusser*	30x30	67.16	67.31	74.68

Table 5: Percentages of correctly classified instances. Pattern spectra were not normalized and components with area $>=4$ were processed.

that is absent in projected galaxies. Normalizing the pattern spectra values seems to bring little benefit. The main differences are due to the smaller feature space having a lower number of bins and to the area of the components considered. A larger decrease of correct classifications was expected with the normalization of the pattern spectra than the one observed. In principle, normalization would make the method to account less for the existing size differences present in some images of the dataset between mergers and overlapping galaxies. However, using scale-invariant moment invariants can have counteracted this effect. Fig. 3a and Fig. 3b show two cases of correctly classified mergers and projected galaxies, respectively. Fig. 3c shows the large merger Messier 49. Its companion galaxy is a dwarf irregular galaxy, cropped out of the picture. In this figure, it is interesting to notice that the smaller galaxy is correctly classified as overlapping, in spite of being also visually linked to the wide halo. Fig. 3d shows instead two misclassified galaxies: both the galaxies are classified as mergers, but they are actually overlapping. Surely, the small area of the objects in Fig. 3d does not help the classification. Fig. 4a shows two (or more) merging galaxies at a very late stage. It is very difficult to define the boundaries of the galaxies involved and the image looks like

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

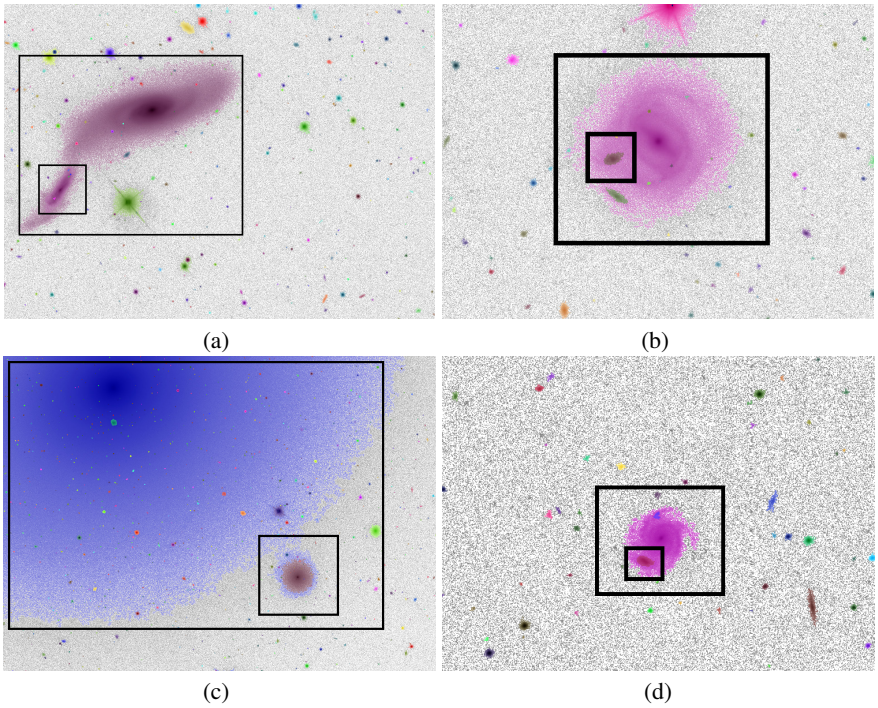
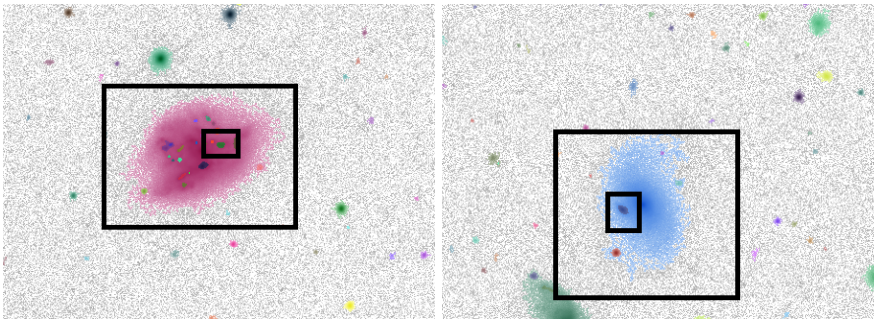


Fig. 3: (a) and (b) show correctly classified merging and projected galaxies, respectively; in (c) the small galaxy is correctly classified as overlapping; in (d) the two galaxies are overlapping but they are classified as mergers. Separate objects are segmented in different colours. The black rectangles highlight which objects we are referring to.

over-segmented. The larger galaxy is correctly classified as a merger, but the companion is not clearly defined. The same happens for the overlapping galaxies in Fig. 4b. The large one is correctly classified, the smaller one is not. In general, it is of course difficult to classify objects that span a few pixels. The classification could be possibly improved with a better separation among close-by objects. We noticed that, during the segmentation, it can happen that too many pixels are assigned to the larger objects, thus reducing the amount of available information for the smaller galaxies. In a previous thesis work [19] compiled at the Kapteyn Astronomical Institute of Groningen, multi-scale connectivity [4] was used to create feature vectors made of collection of 1D *size* pattern spectra at several scales. Such pattern spectra were not local: a single pattern spectrum was produced from every image and object segmentation was not performed. Running the same algorithm of [19] on our dataset gave a percentage of correctly classified objects of 77.81% with C4.5 tree classifier with bagging. The main issue with this solution was that several (hundreds) astronomical objects are present in every image: it is not very clear what is being classified if objects are not segmented and global pattern spectra are used. On the contrary, classification based on segmented objects guarantees

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.



(a) (b)

Fig. 4: (a) last stage of a merging phase with small segmented structures and (b) a small projected galaxy made of a few pixels. In both cases, classification is a difficult task. Separate objects are segmented in different colours. The black rectangles highlight which objects we are referring to.

a more reliable and truthful outcome and evaluation of the results. We are not aware of other approaches that try to distinguish mergers from overlapping galaxies, since the focus of most of the papers is rather on classifying the different morphologies.

The parallel algorithm to compute local pattern spectra was tested for speed performance. It was implemented in C language with POSIX Threads. A shared-memory Dell R815 Rack Server with four 16-core AMD Opteron processors and RAM memory of 512GB was used in the tests. Table 6 at row 1 and 3 shows the run-time and the speed-up value to compute the pattern spectra of hundreds of stars and galaxies in each of the 98 images, with 30x30 pattern spectra. Flusser's and Hu's moments were used. Run-time decreases from almost 7 minutes to 1 minute and a half on 8 threads. In general, the images in the SDSS are small, about 3Mpx resolution and the performance is affected more by the time spent merging the results of the different threads than by the time spent in the actual computation of moments and pattern spectra. Moreover, load balance is not optimal: in case the partition assigned to a thread contains a small number of components or any object at all, the thread will have a small computational work. We decided then to test the parallel computation of 2D local pattern spectra on a large radio cube with 360x360x1464 resolution, named WSRT (Westerbork Synthesis Radio Telescope, courtesy of P. Serra). It contains radio emission values from galactic sources. The results are shown in Table 6, at row 2 and 4. The time to compute the pattern spectra for the objects identified (about five hundred) goes from 4 minutes on a single threads to 34 seconds on 32 threads. Load balance issues are still evident in the speed-up computation, though. We recall here that an important step of the pipeline that leads to object classification is also the segmentation of astronomical objects. In the case of the WSRT cube, segmentation was done in parallel, as in [13]: run-times went from 11 minutes on single thread to 2 minutes and 10 seconds on 16 threads, on the same Dell machine.

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

Test / Threads:	1	2	4	8	16	32	64
Run-time SDSS (s)	421.47	227.09	130.32	91.53	104.05	136.62	258.28
Run-time WSRT (s)	242.13	156.91	100.01	46.15	40.63	34.81	46.18
Speed-up SDSS	1.00	1.86	3.23	4.60	4.05	3.08	1.63
Speed-up WSRT	1.00	1.54	2.42	5.25	5.96	6.96	5.24

Table 6: Execution times (in seconds) and speed-up values obtained after computing in parallel 2D local pattern spectra, for the 98 SDSS images and the WSRT cube, with 30x30 pattern spectra for each object in the images.

7 Conclusions and future work

The use of collections of local 2D pattern spectra as feature vectors suitable for classification of astronomical object looks promising. Experiments were made on a dataset of 196 galaxies from the SDSS Data Release 7. The goal was to classify if two close-by galaxies present in each image were either merging or overlapping. The galaxies were automatically segmented by our own segmentation algorithm. A set of local 2D pattern spectra, binned using the size of the components in the segmented galaxies and image moment invariant information is computed in parallel. The Weka C4.5 tree classifier with bagging gave the best classification results: the percentage of correctly classified instances is about 80%. In future work, other attributes could be investigated, for example shape measures derived from moments. Other classifiers should also be tested. Neural networks approaches as in [2] look promising. Further improvements to object segmentation following more accurately the brightness profiles could possibly enrich the quality of the pattern spectra and improve classification.

References

1. Abazajian, K.N., Adelman-McCarthy, J.K., Agüeros, M.A., Allam, S.S., Allende Prieto, C., An, D., Anderson, K.S.J., Anderson, S.F., Annis, J., Bahcall, N.A., et al.: The Seventh Data Release of the Sloan Digital Sky Survey. The Astrophysical Journal Supplement Series 182, 543 (Jun 2009)
2. Banerji, M., Lahav, O., Lintott, C.J., Abdalla, F.B., Schawinski, K., Bamford, S.P., Andreescu, D., Murray, P., Raddick, M.J., Slosar, A., Szalay, A., Thomas, D., Vandenberg, J.: Galaxy zoo: reproducing galaxy morphologies via machine learning. Monthly Notices of the Royal Astronomical Society 406(1), 342–353 (2010)
3. Bertin, E., Arnouts, B.: Sextractor: software for source extraction. Astronomy and Astrophysics, Suppl. Ser 117, 393–404 (1996)
4. Braga-Neto, U., Goutsias, J.: A multiscale approach to connectivity. Comp. Vis. Image Understand. 89, 70–107 (2003)
5. Breen, E.J., Jones, R.: Attribute openings, thinnings and granulometries. Comp. Vis. Image Understand. 64(3), 377–389 (1996)
6. Flusser, J., Suk, T.: Construction of complete and independent systems of rotation moment invariants. In: Petkov, N., Westenberg, M.A. (eds.) Proc. Comput. Anal. Images Patterns 2003. Lecture Notes in Computer Science, vol. 2756, pp. 41–48. Groningen, The Netherlands (August 25-27 2003)

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.

7. Flusser, J.: On the independence of rotation moment invariants. Pattern Recognition (PR) 33(9), 1405–1410 (2000)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009), http://doi.acm.org/10.1145/1656274.1656278
9. Hu, M.K.: Visual pattern recognition by moment invariants. Information Theory, IRE Transactions on 8(2), 179–187 (February 1962)
10. Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J.: Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society 389(3), 1179–1189 (Sep 2008)
11. Lotz, J.M., Primack, J., Madau, P.: A new nonparametric approach to galaxy morphological classification. The Astronomical Journal 128(1), 163 (2004)
12. Maragos, P.: Pattern spectrum and multiscale shape representation. IEEE Trans. Pattern Anal. Mach. Intell. 11, 701–715 (1989)
13. Moschini, U., Teeninga, P., Wilkinson, M.H.F., Giese, N., Punzo, D., van der Hulst, J.M., Trager, S.C.: Towards better segmentation of large floating point 3d astronomical data sets: first results. In: Proceedings of the 2014 conference on Big Data from Space (BiDS'14). pp. 232–235. Publications Office of the European Union (2014)
14. Peng, C.Y., Ho, L.C., Impey, C.D., Rix, H.W.: Detailed structural decomposition of galaxy images. The Astronomical Journal 124(1), 266 (2002), http://stacks.iop.org/1538-3881/124/i=1/a=266
15. Pesaresi, M., Wilkinson, M.H.F., Moschini, U., Ouzounis, G.K.: Concurrent computation of connected pattern spectra for very large image information mining. In: ESA-EUSC-JRC 8th Conference on Image Information Mining. pp. 21–25. Publications Office of the European Union (2012)
16. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
17. Salembier, P., Oliveras, A., Garrido, L.: Anti-extensive connected operators for image and sequence processing. IEEE Trans. Image Proc. 7, 555–570 (1998)
18. Serra, J.: Image Analysis and Mathematical Morphology. II: Theoretical Advances. Academic Press, London (1988)
19. Starkenburg, T.: Classifying galaxies with mathematical morphology (2009), https://www.mysciencework.com/publication/read/7470103/classifying-galaxies-with-mathematical-morphology
20. Teeninga, P., Moschini, U., Trager, S.C., Wilkinson, M.H.F.: Improved detection of faint extended astronomical objects through statistical attribute filtering. In: Benediktsson, J.A., Chanussot, J., Najman, L., Talbot, H. (eds.) Mathematical Morphology and Its Applications to Signal and Image Processing. Lecture Notes in Computer Science, vol. 9082, pp. 157–168. Springer International Publishing (2015), http://dx.doi.org/10.1007/978-3-319-18720-4_14
21. Urbach, E.R., Roerdink, J.B.T.M., Wilkinson, M.H.F.: Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. IEEE Trans. Pattern Anal. Mach. Intell. 29(2), 272–285 (Feb 2007)

The final publication is available at link.springer.com.
DOI: 10.1007/978-3-319-23117-4_11.