University of Groningen

Faculty of Science and Engineering

Research Internship

Attribute profiles for galaxy classification

Loran Oosterhaven

supervisors
dr. M.H.F. Wilkinson
A. Nolte

September 2, 2019

Abstract

Astronomical sky surveys contain a large number of images, each containing hundreds of thousands of objects. Automatic extraction and classification of objects is therefore needed. This research project focuses on the derivation of a new kind of feature vector that is suitable for use in a prototype-based classification algorithms. Automatic object detection is performed by converting an astronomical image into a max-tree structure and filtering such a structure. Each object in max-tree representation is traced from the central peak to its faintest detectable regions. Geometric and brightness attributes are sampled at each node to obtain an *attribute profile*, in the likeness of an astronomical brightness profile. The profiles are fit and sampled to extract features, and automatic labelling is achieved by matching a large number of astronomic images against a GalaxyZoo dataset. A classification experiment is performed using a test dataset containing 400 objects to determine feature relevances. The derived feature vector is based on the affine moment invariants and brightness profile. A score of 76.5% is achieved using the localized GMLVQ (LGMLVQ) classifier.

Contents

Chapter 1

Introduction

Telescopes generate thousands of publicly available, high-resolution images containing astronomical objects such as galaxies and stars. These objects have to be detected (i.e: finding which pixels of the image belong to an object and not to the background) and classified, but it is simply too time-consuming and error-prone to perform this task manually. Motivated by the successful application of moment invariants in [20] in a max-tree form to derive a feature vector, and the concept of galaxy brightness profiles, the idea of attribute profiles will be explored in this paper. The idea is that an attribute profile gives information of the attribute (i.e. shape and brightness) of an object at different brightness levels, which correspond to max-tree nodes, similar to an astronomical brightness profile. Shape descriptors of circularity, convexity, and affine moment invariants will be considered. Furthermore, a derivation will be given on how to analytically fit the obtained brightness profile. In order to obtain and label the data for testing and experimentation purposes, a program design is described that uses GalaxyZoo data to label the data automatically. A dataset of 400 galaxies of 4 different classes is obtained and a classification experiment is applied to it. The classes that are used correspond to the Hubble Types E0, E5-E7, SB, Sb.

This paper is structured as following:

1. Chapter 2 will introduce the reader to the necessary mathematical and astronomical theory required to understand the rest of the paper.

2. Chapter 3 will discuss the obtained results, describing how the data is obtained and how the features are extracted and fit. It ends with a classification experiment with an obtained dataset and derivation of the new feature vector.

3. Chapter 4 closes the paper with the conclusions and possible direction for further research.

Chapter 2

Background

2.1 Morphological classification of galaxies

There are several schemes in use by astronomers to classify galaxies according to their morphology (structural properties). The most well-known of these schemes is the Hubble Sequence invented in 1926 by Edwin Hubble [7]. It is often visualized by the famous *tuning fork diagram* (Figure 2.1). The Hubble Sequence divides galaxies into three main lines, based on their most prominent features.

- Elliptical
- Spiral
- Barred spiral

In addition to these main lines the Hubble Sequence admits a fourth category called *Irregulars*, divided into Irr I, which are galaxies that are irregular but show some kind of structure; Irr II, irregulars that are completely disorganized.

Elliptical galaxies (E) look like ellipses and are assigned a *Hubble Type*, $E = 10 \times (1 - b/a)$ where a, b are the respective lengths of the semi-major and semi-minor axis of the ellipse. Observed elliptical galaxies range from E0, which is a perfectly spherical galaxy, to E7, which has the most flat shape.

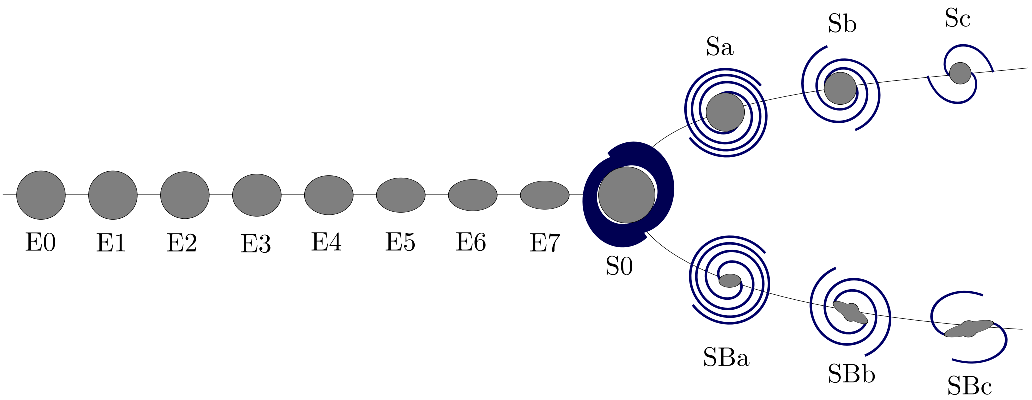
Spiral galaxies (S) consist of a flat, rotating disc containing gas, stars and dust, and a central concentration of stars known as the bulge. Roughly two-thirds of observed spiral galaxies also have a 'bar' extending from the central bulge. Such galaxies are known as barred spirals (SB). Spiral and barred spiral galaxies are further divided into 'late' or 'early' type according to the prominence of their bulges, which is given by the ratio of the luminosity of the bulge and the total luminosity and tightness of the spiral arms. This ratio therefore also provides a straight-forward quantity to be used in automatic classification. The letter following S or SB indicates this, where 'a' means earliest and 'c' means latest. An intermediate class between the S and SB series is the SB0 class, which shows a bulge but no other features like spirals or bars.

More elaborate classification schemes have been proposed by De Vacouleurs [9] and Morgan [10], but these are not within the scope of this project.

2.2 Galaxy brightness profiles

The surface brightness of a galaxy may be described with a function or *profile* $I(r)$ which yields the brightness of the galaxy measured at the boundary of an isophote with a radius r of the same shape as the galaxy. The two most common profiles in use by astronomers are the exponential profile:

$$I(r) = I_0 e^{-b\frac{r}{r_e}}$$



Source: `https://commons.wikimedia.org/wiki/File:Hubble_sequence.svg`

Figure 2.1: Tuning fork diagram of the Hubble Sequence

and the Sérsic profile which was first published by Sérsic in [13]:

$$I(r) = I_0 e^{-b\frac{r}{r_e}^{1/n}}.$$

In both equations I_0 refers to the central brightness, r_e refers to the half-light radius, also known as *effective radius* which is the radius that contains half of the light of the system, and b is a constant depending on n. In the last equation n is known as the Sérsic parameter.

Obviously, the exponential profile is a special case of the Sérsic profile with $n = 1$. The outer disks of spiral galaxies can generally be modelled with an exponential profile and elliptical galaxies, bulges, bars, and psuedo-bulges can generally be modelled with a Sérsic profile with typically $n \approx 4$, which is also known as *De Vacauleur* profile. This means that the profile of a spiral galaxy should be modelled with a combination of two profiles: an exponential profile in the bulge and a Sérsic profile in the outer disk.

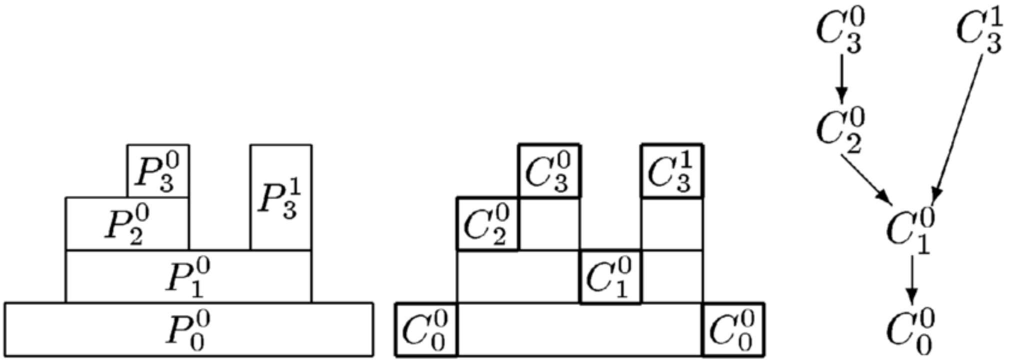
2.3 Max-tree segmentation

Any image can be decomposed to a tree where the nodes contain connected components, i.e pixels that are pathwise connected with the same intensity. A hierarchy of connected components can be obtained by tresholding the image at every intensity level. Since the intensity is ordered, components with higher intensity will be nested on top of components with lower intensity. Such a hierarchical representation of the image is called a max-tree [19]. The leaves in such a tree represent the local maxima of the image and the root represents the entire image with 0 intensity, i.e the black background. The left side of figure 2.2 illustrates this idea of a component hierarchy for a 1D signal. The same principle can be applied to 2D images. The right side of the figure illustrates the parent-child relationships. This tree can also be used to efficiently find interesting attributes of the nodes, such as area and average intensity, or in the case of this project: the geometric moments.

2.4 Background estimation and object detection

In this section we will describe the method used by MTObjects created by Teeninga et. al [12] to estimate the background and detect astronomical sources.

Before max-tree segmentation can take place, the background of the image should be removed so that the intensity values will be approximately 0 on parts of the image where there are no objects. The astronomical images from the SDSS dataset (see section 3.1) are assumed to be a sum of a background image B, noiseless image O and Gaussian noise. Therefore, the background value can be estimated by the mean-value of flat-tiles. A flat-tile is a region of the image devoid



Source: Figure 4 [25]

Figure 2.2: The max-tree structure: each P_h^k is the h'th peak component at threshold level k of a 1D signal. (left); node at each signal value (center); the implied hierarchy between the max-tree nodes (right)

of objects. All possible flat-tiles of the image of size 64x64 pixels are considered, and rejected or accepted based on a K^2 normality test and a t-test of equal means. If a tile is valid its mean value added to a list. After all tiles have been processed the background value B is estimated as the mean value of the list.

After the max-tree is built MTObjects can perform the object detection algorithm. The first step of this algorithm determines which nodes are significant (if it represents one or more objects) using a statistical test, given the background estimate. The second step identifies which nodes belong to which objects and identifies nested objects on top of other larger objects. This operation is called *deblending.* For the specifics we refer the reader to [12].

2.5 Shape analysis

For the problem of classifying images according to shape numerical measures are needed which carry geometric and structural information about the image. Such measures are called *shape descriptors* and are widely used in image and shape classification problems. Each shape descriptor has advantages and disadvantages and its suitability depends on the problem at hand. In this project the moment-based descriptors *affine moment invariants*, *circularity* and a *convexity* measure will be considered.

2.5.1 Moments

Geometric moments, also called raw moments are widely used in pattern recognition tasks. The moment $m_{p,q}$ with order $n = p + q$ of a shape S is given by

$$m_{p,q}(\mathcal{S}) = \iint_S P_x^p P_y^q f(x,y) \tag{2.1}$$

where f is known as the *image function.* In our case the shape $\mathcal{S}$ is a set of pixels, and the integral is replaced by summation. The image function is simply 1 if the point (x, y) is in $\mathcal{S}$, and 0 otherwise. Moments computed with such an image function are called binary image moments.

The raw moments contain basic information about the shape, for example it is easily verified $m_{0,0}$ is equal to the shape's area and $c_x = \frac{m_{0,1}}{m_{0,0}}, c_y = \frac{m_{1,0}}{m_{0,0}}$ are the x, y coordinates of the centroid of $\mathcal{S}$. Central moments are moments that have been normalized w.r.t translation and scaling, by moving the origin to the shape's centroid and dividing by the factor $m_{0,0}$. They can be computed from raw moments as:

$$u_{p,q} = \sum_{m=0}^{p}\sum_{n=0}^{q}\binom{p}{m}\binom{q}{n}(-1)^{m+n} \cdot x_c^m y_c^n \cdot m_{p-m,q-n} \tag{2.2}$$

Where c_x, c_y is the centroid of S as defined earlier. This is the form that is used in this project.

Raw moments can easily be calculated incrementally for each node while the max-tree is being built. The initial value for each node is simply $m_{p,q}(x, y) = x^p y^q$, and when two nodes are merged, the moments of the resulting nodes are computed as the sum of the nodes being merged.

2.5.2 Image moment invariants

Over the last decade moment invariants have become a very popular tool for shape recognition. They were first introduced by Hu [1] who derived a set of 7 independent second and third order invariants. These moment invariants are computed from image moments and they have the property of being invariant to scale, translation, and rotation. However, Hu's set was proven to be algebraically dependent. More recently, Flusser and Suk [2] derived a set of 4 *affine moment invariants*, which are invariant under a general affine transformation of the image, including reflection and non-uniform scaling. Such invariance is preferable when considering galaxies as reflection should not affect the galaxy class. This set is also proven to be algebraically independent and complete, which means that no invariant in the set can be computed from an algebraic expression of the other moments and no other invariants of the same order exists. This property is desirable for image classification tasks.

The construction in [2] starts with the definition of complex moment $c_{p,q}$ of order $(p+q)$ which is defined as

$$c_{p,q} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x + iy)^p(x - iy)^q f(x, y)dxdy. \tag{2.3}$$

Each complex moment can be expressed in terms of raw moment $m_{p,q}$ as

$$c_{p,q} = \sum_{k=0}^{p}\sum_{j=0}^{q}\binom{p}{k}\binom{q}{j}(-1)^{q-j} \cdot i^{p+q-k-j} \cdot m_{k+j,p+q-k-j}. \tag{2.4}$$

The set of affine moment invariants up to third order is computed from the complex moments as

$$\begin{aligned} A_1 &= c_{11} & A_3 &= c_{21}c_{12} \\ A_2 &= c_{30}c_{03} & A_4 &= \Re(c_{30}c_{12}^2). \end{aligned} \tag{2.5}$$

2.5.3 Convexity

Convexity can be considered a basic property of shape. A planar shape $\mathcal{S}$ is said to be convex if it has the following property: if point A and B belong to S then all points from the line segment AB belong to S as well. The convex hull of S is denoted $CH(\mathcal{S})$, which is the smallest convex shape that contains S. This leads to a straight-forward definition of convexity [14]:

$$\mathcal{C}(\mathcal{S}) = \frac{Area(\mathcal{S})}{Area(CH(\mathcal{S}))} \tag{2.6}$$

This measure of convexity has the following properties:

- $\mathcal{C}(\mathcal{S}) \in (0, 1]$.
- $\mathcal{C}(\mathcal{S}) = 1$ iff $\mathcal{S}$ is convex.
- there are shapes whose convexity is arbitrarily close to 0.
- $\mathcal{C}(\mathcal{S})$ is invariant under similarity transformations.

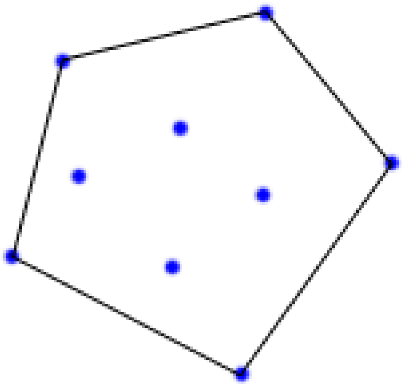


Figure 2.3: The convex hull CH of a point set $\mathcal{S}$

- $\mathcal{C}(\mathcal{S})$ is robust w.r.t to small changes in area (noise).

Another convexity measure is based on generating points from a set and measuring the probability that a point dividing the the corresponding line segments belong to the same set [22]; yet another is defined similarly to equation (2.6) but using shape perimeter. Since a max-tree does not directly provide access to perimeter, but does store an area attribute for each node, (2.6) provides the best option for this project. It is expected that elliptical galaxies should display a higher measure of convexity than non-elliptical galaxies in the centre/bulge/disc regions.

2.5.4 Circularity

Circularity, sometimes referred to as *compactness* is another shape descriptor which defines the degree to which a shape deviates from a circle. Žunić et. al defined a circularity measure calculated from central moments [15]:

$$\mathcal{K}(\mathcal{S}) = \frac{(u_{0,0})^2}{2\pi(u_{2,0} + u_{0,2})} \tag{2.7}$$

$\mathcal{K}(S)$:This measure can be shown to have the following properties:

- $\mathcal{K}(\mathcal{S}) \in (0, 1]$.
- $\mathcal{K}(\mathcal{S}) = 1$ iff the S is a circle.
- there are shapes whose measured circularity is arbitrarily close to 0.
- $\mathcal{K}(\mathcal{S})$ is invariant under similarity transformations.

2.6 Classification

A popular automated classification scheme is as Learning Vector Quantization (LVQ) as introduced by Kohonen [21]. The benefits of this algorithm include its relatively low complexity; its ability to naturally deal with multi-class problems and the ease with which the results can be intuitively interpreted by a human.

An LVQ model consists of a number of prototypes $W = \{\mathbf{w}_i, c(\mathbf{w}_i)\}$, where $\mathbf{w}_i \in \mathbb{R}^n$ is a prototype, represented by a location in a real number space of dimension n and $\mathbf{c}(\mathbf{w_i})$ is the class of the prototype. Note that each class may be associated with multiple prototypes. In the learning (also known as training) phase, the prototypes are placed in the optimal position given the input data. In the original version of LVQ a heuristic algorithm was used, but in 1996 Sato & Yamada [8] reformulated the problem by optimization of the cost function

$$\sum_i \phi(\mu_i) \quad \text{where } \mu_i = \frac{d_J^\lambda(\mathbf{x_i}) - d_K^\lambda(\mathbf{x_i})}{d_J^\lambda(\mathbf{x_i}) + d_K^\lambda(\mathbf{x_i})} \tag{2.8}$$

based on the steepest descent method. $d_J^\lambda(\mathbf{x}_i) = d^\lambda(\mathbf{w}_J, \mathbf{x}_i)$ is the distance of a data point $\mathbf{x}_i$ to the closest prototype $\mathbf{w}_J$ with the same class label y, and $d_K^\lambda(\mathbf{x}_i) = d^\lambda(\mathbf{w}_K, \mathbf{x}_i)$ is the distance of a data point $\mathbf{x}_i$ to the closest prototype $\mathbf{w}_K$ with a class label different from y. After the model has been trained, an input data point $\mathbf{x}_i \in \mathbb{R}^n$ can be classified by finding the class of the nearest prototype.

LVQ has a number of drawbacks which include its reliance on a euclidean metric and its inability to account for correlation among features in the input data. Therefore improvements to LVQ have been proposed. One of these improved schemes is known as Generalized Matrix LVQ or GMLVQ [23]. This algorithm replaces the euclidean metric with a quadratic form with matrix $\Lambda = \Omega^T\Omega$ also known as the *relevance matrix* as it gives the relevance factors for each input dimension and also pairwise correlations. Λ is also learned (optimized) along with the prototypes in the learning phase. A variant of GMLVQ where each prototype learns its own metric is called localized GMLVQ or LGMLVQ.

Chapter 3

Results

3.1 Program design

Astronomical images from the SDSS (Sloan Digital Sky Survey) come in the `FITS` file format. These are high resolution grayscale images with up to 32 bits of precision per pixel, defined on a particular color band, which is one of u, g, r, i, z.

Over the course of the research project a program was developed that extracts attribute profile features from astronomical images and performs the labelling for classification purposes automatically. As a basis for this program the python library `mtobjects` is used which is freely available [11] and provides an implementation of the max-tree algorithm, background estimation and object detection discussed in sections 2.3 and 2.4 respectively. The source code was modified in order to have it compute the matrix of geometric moments up to 4th order while the max-tree is being built. This means that each node is augmented with a 4x4 matrix. A move factor (see [4]) parameter of 1.5 was used, which tells `mtobjects` to include less of the faint outskirts of objects.

After background estimation and subtraction, and the max-tree is built, the next step in the pipeline is to process each object and traverse its nodes from the peak down to the faintest detectable nodes. The node with the highest brightness (peak) is found, and this is used to compute the object's coordinates. If a match with a GalaxyZoo(GZ) object is found, i.e: the coordinates are sufficiently close to an object in the GZ dataset, the label of the object is known the program continues to traverse the node's parents until it either meets the root node of the image, or a node is reached that does not belong to the object. While this traversing occurs, at each step a list of the following attributes is maintained:

1. Radius r, defined as $\sqrt{\texttt{Area}}$
2. Brightness $I(r)$
3. Affine moment invariants, derived from the raw moments calculated in each node using (2.5). The quantities are rescaled as follows, to bring them all in a similar range:

$$\begin{aligned} A'_1 &= 50A_1 & A'_3 &= 5 \cdot 10^3 A_3 \\ A'_2 &= 10^5 A_2 & A'_4 &= 10^8 A_4 \end{aligned} \tag{3.1}$$

4. Circularity, $\mathcal{K}(r)$ is calculated from the raw moments using 2.7. If the denominator equals zero ($u_{2,0} + u_{0,2} = 0$), the formula becomes invalid. In this case a value of $\mathcal{K}(r) = 1$ is used.

Thus, after the traversing is finished, the shape properties are obtained as function of radius r. In addition to these quantities also convexity is calculated as a function of brightness. This is not done by max-tree traversal, but rather by first computing a list of pixels sorted by brightness and iterating over each pixel (x, y) in increasing order. Each iteration the corners of the pixel, i.e: $\{(x, y), (x + 1, y), (x, y + 1), (x + 1, y + 1)\}$ are inserted into a point set and convexity is calculated using (2.6) by (re)computing the convex hull. As an optimization, the convex hull is only computed a maximum of 150 times on evenly spaced intervals of brightness, because some objects contains tens of thousands of pixels, an unreasonable amount of time would be spent recomputing convex hulls computations. The reason for not doing the computation during max-tree traversal is that there is not an easy way to query which pixels belong to each node using `mtobjects`. This functionality is certainly possible but due to time constraints it was omitted in this project.

A plot of each function along with a relevant image cutout is output in a folder structure organized by galaxy type. These plots will be used to perform an initial visual exploratory analysis of the data.

3.2 Data acquisition and automatic labelling

In order to provide the program with test data, i.e `FITS` files and labels of the objects a structured method is needed. The SDSS contains hundreds of thousands of such files, but only rarely is it possible to accurately determine a classification for a given object by visual inspection especially if an object is composed of only a handful of pixels.

GalaxyZoo 2 [6] is a citizen science project that provides classifications for hundreds of thousands of galaxies. Any person can contribute to the project by using a web tool where the user has to answer a number of questions about a photo of a particular astronomical object. For this research project a smaller dataset is preferable so table 8 of the GZ2 classification data was used, containing 19,765 galaxies classified in the coadded (runs 106 and 206) Stripe 82 images. This list is referred to as L. The data in L is provided in `csv` file format, where each row includes the equatorial coordinates α, δ (right ascensions and declination) and a shorthand string that describes the most probable classification. The format of this shorthand string is described in the appendix of [18]. Given L, the SDSS Science Archive Server [17] was queried to retrieve the list of corresponding `FITS` images (g-band) for each sky coordinate. The program was also modified to accept the file containing L as input specified by a program argument. If the program is supplied with such a file, it will look up the closest object (α, δ, label) in the list by computing the distance of the coordinates of the object's peak to each object's coordinates. To compute the peak's coordinates from it's pixel coordinates, the `FITS` header fields `CRPIX1, CRPIX2, CRVAL1, CRVAL2, CD1_1, CD1_2, CD2_1, CD2_2` are used. These fields are extracted on program initialization and combined in a matrix. To compute the sky coordinates from the pixel coordinates given the pixel coordinates of the peak node `peak_x, peak_y` the following equation is used:

$$\begin{pmatrix}\alpha\\ \delta\end{pmatrix} = \begin{pmatrix}\texttt{CD1\_1} & \texttt{CD1\_2}\\ \texttt{CD2\_1} & \texttt{CD2\_2}\end{pmatrix}\begin{pmatrix}\texttt{CRPIX1 - peak\_x}\\ \texttt{CRPIX2 - peak\_y}\end{pmatrix} + \begin{pmatrix}\texttt{CRVAL1}\\ \texttt{CRVAL2}\end{pmatrix} \tag{3.2}$$

The shorthand string obtained from the closest match is used as the classification. If the closest distance is greater than 0.05 arc minutes, the program decides no match was found and the algorithm does no further processing on the object. The script `get_all_stripe82.sh` in the source code implements this method.

Refer to listing 3.1 for a brief overview in psuedo-code of how the complete program works, implementing the design discussed in the previous sections.

3.3 Exploratory data analysis

To determine relevant features for classification a brief exploratory analysis is performed on the data. The analysis was carried out as follows: the plots from 3 objects of the GZ2 classes 'Er', 'Ec', 'SBa', 'SBc', 'Sa', 'Sc', 'A', were inspected for general patterns in the data and features that might distinguish each particular galaxy class, where Er and Ec stand for Hubble Type E0-E1, E6-E7 respectively and 'A' stands for stars. The classes were chosen as such to have a large range of galaxy types. Unfortunately for the SBc type only 2 galaxies were available, as no more had

Figure 3.1 Program psuedo-code

```
const gz_catalogue $\leftarrow$ load_GalaxyZoo_Catalogue(gz_filename)
const image_url_list $\leftarrow$ load_urls(url_filename)
for each url $\in$ image_url_list do
    image $\leftarrow$ Download(url)
    test
    mt $\leftarrow$ BuildMaxTree(image)
    for each object $\in$ mt do
        peak_node $\leftarrow$ find_Peak(object)
        peak_coords $\leftarrow$ node_to_skycoords(image.coord_matrix, peak_node)
        distance, label $\leftarrow$ find_match(peak_coords, gz_catalogue)
        if distance $> 0.05$ arcmin then
            continue
        end if
        Convexity $\leftarrow$ convexity(sort_by_brightness(object.pixel_indices))
        R, AMIs, Brightness, Circularity $\leftarrow$ []
        cur_node $\leftarrow$ peak_node
        while cur_node $\in$ object do
            append to R, AMIs, Brightness, Circularity using cur_node attributes
            cur_node $\leftarrow$ parent(cur_node)
        end while
        output_plots(label, R, Brightness, AMIs, Circularity, Convexity)
        output_features(label, R, Brightness, AMIs, Circularity, Convexity)
    end for
end for
```

been encountered by the program at the point of analysis. The data can be found in the eda folder of the code repository.

The first plot that was considered is the logarithm of brightness as a function of radius (log $I(r)$. From first inspection all objects display a profile that can be fit by combination of three piecewise continuous Sérsic profiles. The radius of a peak component is not the same thing as the radius of an isophote, therefore the usual Sérsic profile is not observed even on elliptical galaxies. Another explanation for this piecewise curve could be that it corresponds to different structural parts of the galaxy. Nevertheless, a good fit can be made, as discussed in the next section. Furthermore, it is observed that Er and Ec galaxies display similar brightness profiles. For spiral galaxies the profiles range from looking similar to E type galaxies to a 'sigmoid' shaped profile. SB-type can display this shape even more pronounced. Refer to figure 3.4 for an illustration.

Inferring patterns from inspecting the Affine Moment Invariants (referred to as *AMI*s from now on) plots is more difficult as they are very sensitive to small changes in the object and the profile displays more complex patterns. The meaning of one given moment invariant at a given radius is unclear, as the set of invariants should be viewed together as a description of the shape of an object, rather than in isolation. What can be seen is that:

1. For the Er class, the 3d and 4th AMI is close to 0 throughout the profile.

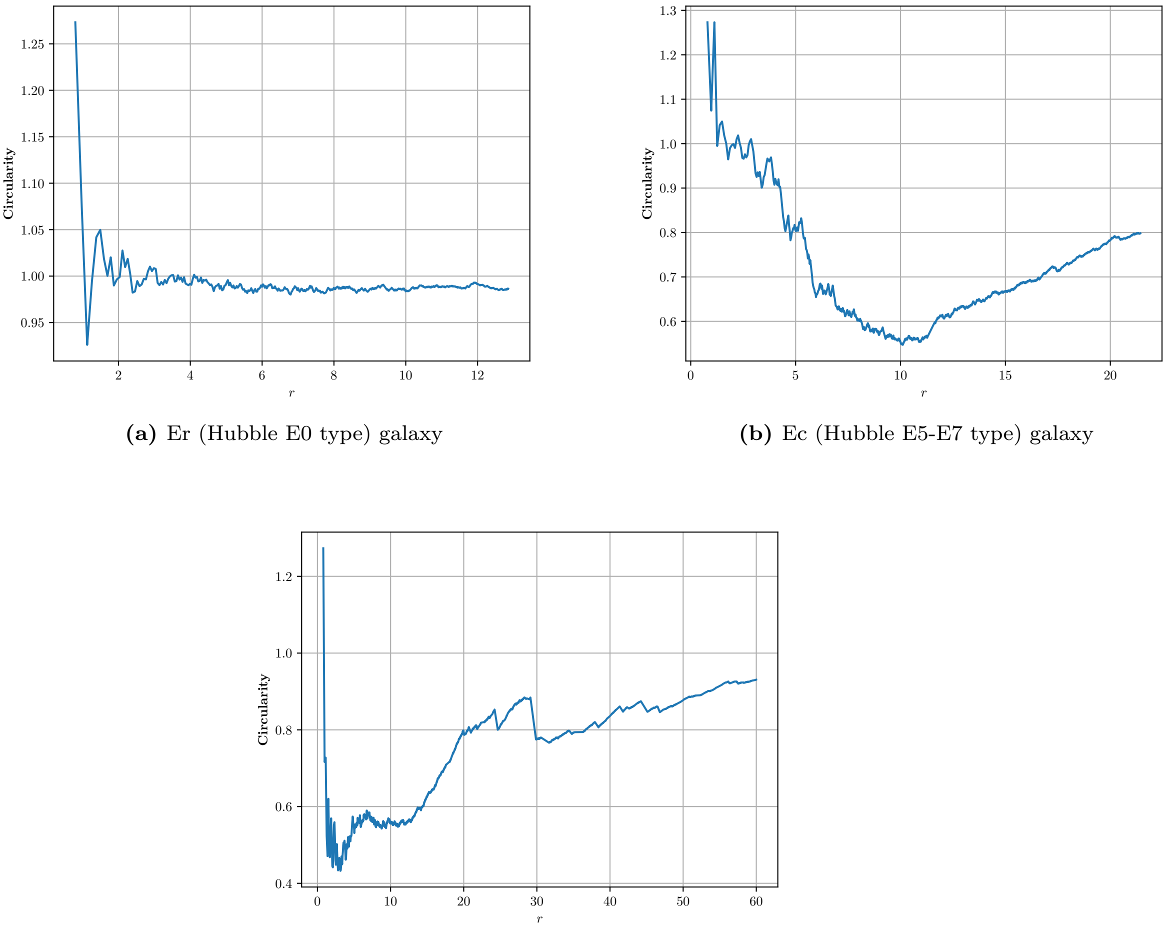
2. The 3d and 4th AMIs tend to be close together in general.

3. For all classes, the value of the first AMI tends to be below or above 1.

4. The barred spirals (SBa, SBc) tend to have larger spikes in the invariants.

Inspecting the circularity plots (see Figure 3.2 for an example), it is immediately obvious that the value quickly converges to 1 for the Er class, while the Ec class remains between $0.6 - 0.8$ throughout the profile. This is reasonable, since Er resembles a circle and Ec have much higher eccentricity. Barred spirals diplay a sharp drop in the early parts of the profile followed by a gradual climb, but for regular spirals no obvious patterns are apparent.

Figure 3.3 shows a few example convexity profiles. The convexity profiles of the elliptical class are rather similar to each other and look roughly quadratic in the radius, although with noise. The initial value is always 1 since $K(\mathcal{S})$ evaluates to 1 for a single pixel. The maximal (final) value being reached is roughly between 0.8 and 1. For spiral galaxies, especially barred spirals the maximal value is lower, and can be as small as 0.6. Also, the profile rises much slower in the central/disc regions. This can be explained by the presence of features such as spirals, especially if they are more loosely winded.

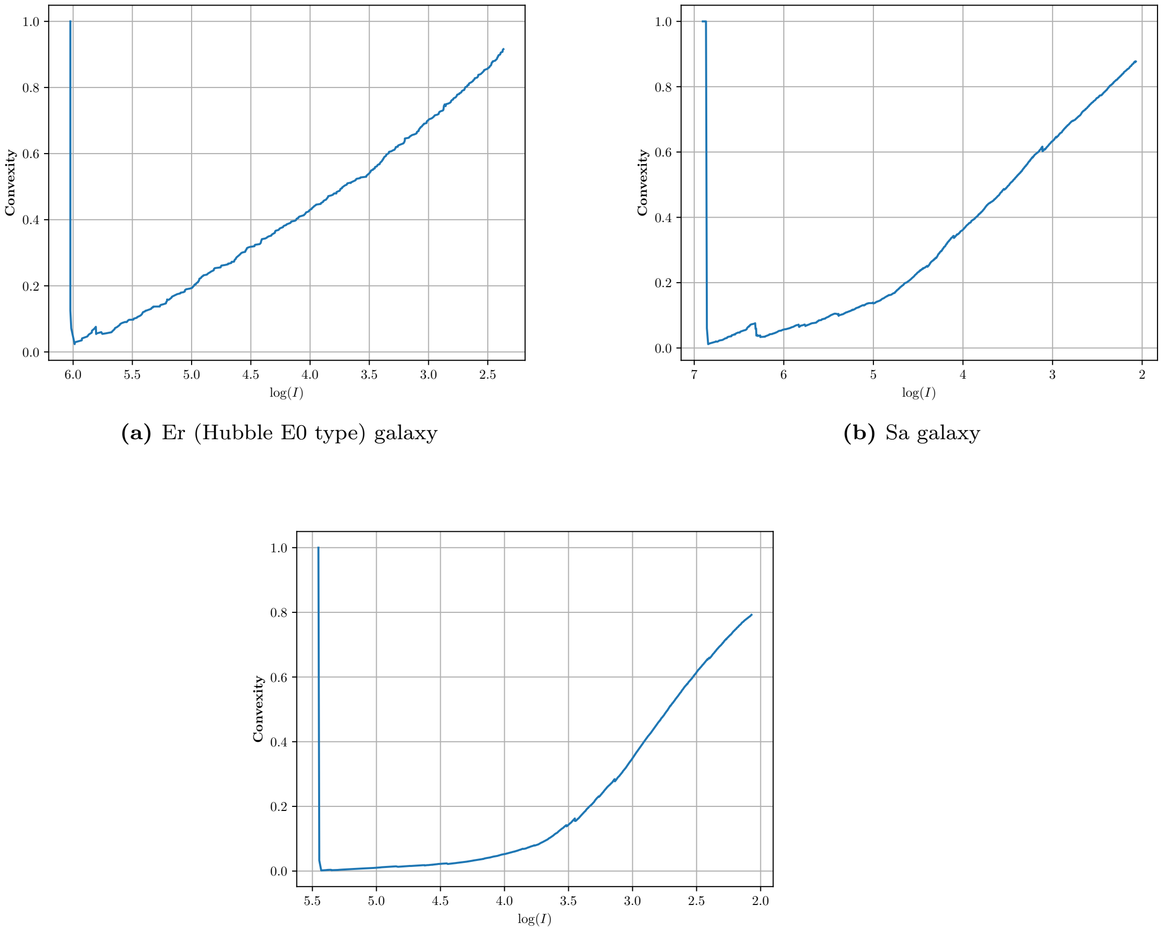


(a) Er (Hubble E0 type) galaxy

(b) Ec (Hubble E5-E7 type) galaxy

(c) SBc type galaxy

Figure 3.2: Circularity profiles.



(a) Er (Hubble E0 type) galaxy

(b) Sa galaxy

(c) SBc type galaxy

Figure 3.3: Convexity profiles.

3.4 Fitting the brightness profile

To fit the object's brightness profile, the following model was used. The profile is modelled as a combination of piece-wise continuous functions defined on three subregions of the profile of the form:

$$\log I(R) = \log I_c - aR^b \quad \text{where } a > 0. \tag{3.3}$$

This corresponds to a Sérsic profile after logarithmic scaling, with $n = 1/b$. For each curve, the values of $\log I_c$, a and b are determined analytically using three data points in the desired range: R_0, R_1, R_2 and their corresponding intensities. For practical and efficiency reasons this is better than having to fit the functions using a fitting algorithm, since it provides a closed-form mathematical expression.

After shifting the origin a distance of R_0 to the right (i.e: $R' = R - R_0$) the following equations are obtained for the three points:

$$\log I(R_0) = \log I_c \tag{3.4}$$

$$\log I(R_1) - \log I_c = -a(R_1 - R_0)^b \tag{3.5}$$

$$\log I(R_2) - \log I_c = -a(R_2 - R_0)^b \tag{3.6}$$

(3.4) directly gives $\log I_c$. The solution for b and a is derived as follows:

$$\frac{\log I(R_1) - \log I(R_0)}{\log I(R_2) - \log I(R_0)} = \frac{(R_1 - R_0)^b}{(R_2 - R_0)^b}$$

$$b = \log_{\frac{R_1 - R_0}{R_2 - R_0}} \frac{\log I(R_1) - \log I(R_0)}{\log I(R_2) - \log I(R_0)}$$

$$a = -[\log I(R_1) - \log I(R_0)]/(R_1 - R_0)^b.$$

Since the coordinate system was shifted before this derivation this shift has to be undone to match the real profile. The profile that is actually obtained using this method is:

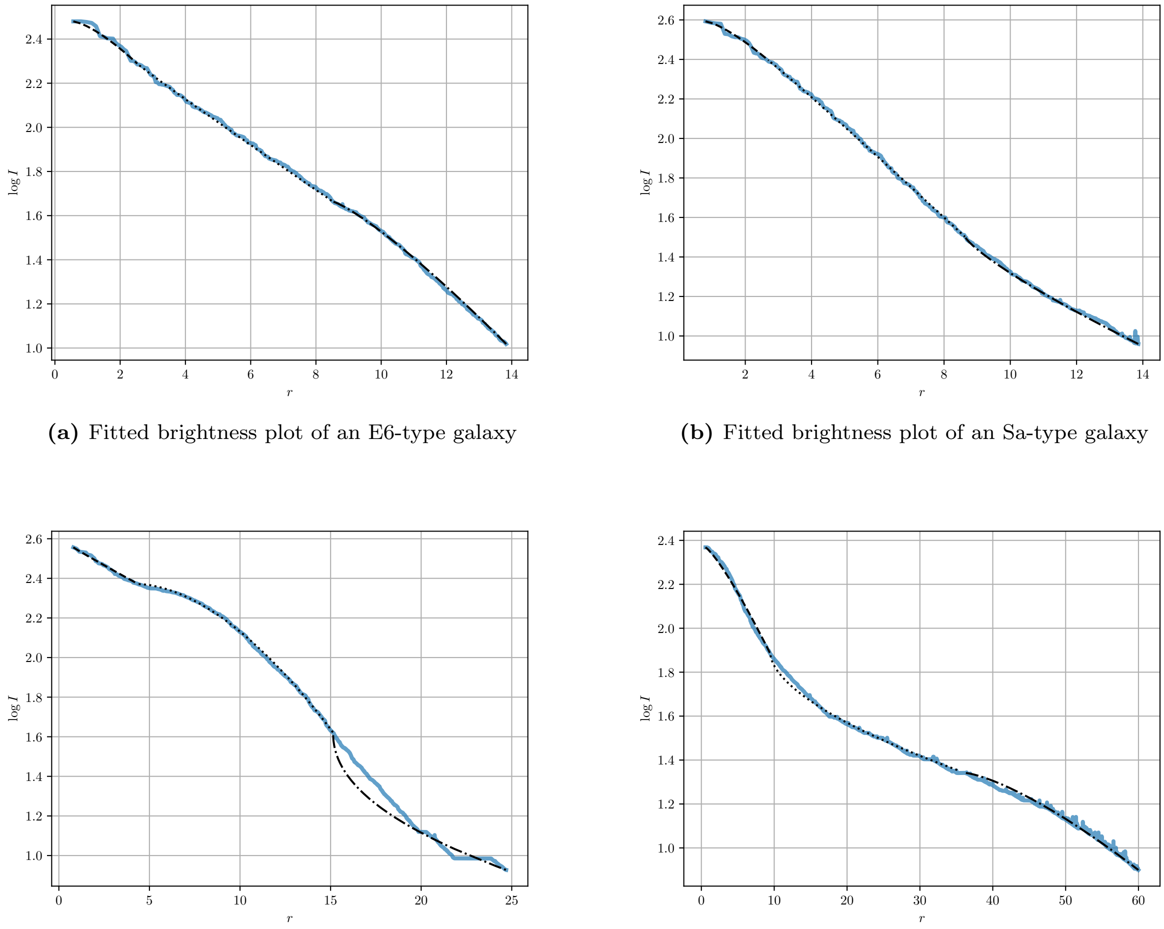
$$\log I(R) = \log I_c - a(R + R_0)^b.$$

It is possible to bring this in the form of (3.3), but this is non-trivial and involves a series expansion.

After some experimentation the profiles were decided to be defined on the following subregions. The values of R_n are computed by linear interpolation between the minimum and maximum radius, i.e: $R_n = (1 - t_n) \times R_{min} + t_n \times R_{max}$

	t_1	t_2	t_3
Region 1	0	0.075	0.15
Region 2	0.15	0.375	0.6
Region 3	0.6	0.8	1

These regions correspond to the inner core/bulge, outer disc / bar / ellipsoid, and outermost faintest regions respectively. After plotting the profiles quite a good result is obtained that works well on many different types of galaxies, see figure 3.4. The fixed regions sometimes do not match the real profile, this could be improved on, see Chapter 4.



(a) Fitted brightness plot of an E6-type galaxy

(b) Fitted brightness plot of an Sa-type galaxy

(c) Fitted brightness plot of an Sc-type galaxy

(d) Fitted brightness plot of an SB-type galaxy

Figure 3.4: Fit brightness plots for four different types of galaxies. The dashed line is the first region, the dotted line is the second region and the dash-dotted line is the third region. Note that the faintest regions of the Sc type were not fit adequately due to the wrong regions being used.

3.5 Feature extraction

To carry out a classification experiment and determine which features are important, features must be extracted from the data that give a good characterization. The most straight-forward thing to do is linear sampling of the value of each attribute at equal intervals. Furthermore, Moschini et. al mentioned in their research [20] that they had success with log-binned area in their local spectra features as nodes with lower area contain more detail. Taking inspiration from this, something similar can be done by using a logarithmic radius scale and taking samples at equal intervals rather than a linear radius. The following features were extracted:

1. The parameters of the fit brightness profiles (3.3) after normalization such that the brightness lies in the $[0, 1]$ range. Therefore, $I_0 = 1$ always for the first region and it is not used. These features are named as $\{$`ea, fy0, fa, gy0, ga, eb, fb, gb`$\}$ in the output parameters file. Corresponding to (3.3), `ea, fa, ga` are the coefficients a, `fy0, gy0` is $\log I_c$ and `eb, fb, gb` are b.

2. 2x10 circularity samples from minimum to maximum radius, linear and logarithmic scale: {`circ0.0, circ0.11, circ0.22,.. circ1.0`} and {`circ_log_0.0, circ_log_0.11, circ_log_0.22,.. circ_log_1.0`}.

3. 10 convexity samples from minimum to maximum brightness, linear scale: {`conv0.0, conv0.11, conv0.22,.. conv1.0`}

4. 10 affine moment invariants samples for each invariant, from minimum to maximum radius, linear and logarithmic scale: {`ami_n_0.0, ami_n_0.11, ami_n_0.22,.. ami_n_1.0`} and {`ami_log_n_0.0, ami_log_n_0.11, ami_log_n_0.22,.. ami_log_n_1.0`} for `n` $= 1 \ldots 4$

A total of $8 + 2 \times 20 + 10 + 80 = 138$ features were extracted.

3.6 Classification experiment

Features were extracted from roughy half of all objects of GalaxyZoo2's Stripe 82 Coadded (Table 8) dataset. From this dataset a subset of 400 objects was chosen divided into the following 4 classes of size 100 each:

- Er, consisting of objects of GZ2 label `'Er'`. This corresponds to a Hubble type of E0.
- Ec, consisting of objects of GZ2 label `'Ec'`. This corresponds to a hubble type of E5-E7.
- SB, consisting of objects of GZ2 label `'SBa2l', 'SBb2l', 'SBc2l'`. These are SB galaxies with 2 arms and a loose winding.
- Sb, consisting of objects of GZ2 label `'Sb2m', 'Sb', 'Sb2l'`. These are Sb galaxies with medium and loose winding.

Seven different sets of features were analyzed, containing the fit brightness data, circularity at linear and logarithmic scales, convexity, AMI data at linear and logarithmic scale respectively and finally, a combined feature vector containing all of the data. These features are the same as described in the previous section. Note that the first of each 10 samples is not used in the analysis as it comes from either a single pixel or a very small node. This first sample was removed from the program as it is not needed, however the feature extraction script was already running for days before this was discovered so it was too late to fix. Both the GMLVQ and LGMLVQ methods defined in section 2.6 were applied using an open-source implementation: `sklearn-lvq` [24]. Model validation was done using 10-fold cross validation and a regularization parameter of 0 was used. The number of prototypes per class was 1. The classification scores can be found in Tables 3.1 and 3.2, respectively. The generated relevance profiles can be found in Appendix A. The generated confusion matrices can be found in Figures 3.6 and 3.7.

The set of AMI features taken from linear samples gave a performance of around 71% in GMLVQ and 64.25% in LGMLVQ. Furthermore both feature vectors composed of circularity features did well at around 62% in GMLVQ and 70% in LGMLVQ. The performance of the brightness features was in the 45-50% range, but performed above average in distinguishing between Er and Ec types. Somewhat surprisingly, the convexity feature vector did not perform well with 34% in GMLVQ and 31.5% in LGMLVQ. All feature vectors performed better at classifying the Er and Ec classes than SB and Sb classes. A possible explanation is that the data from these classes comes from different underlying GZ2 classes as mentioned in the start of this section. Logarithmic sampling of the AMI and circularity features did not improve or even worsen the classification performance. This suggests that in general, samples from the entire profile are important.

Among the AMI features the first, second and third affine invariants A_1, A_2, A_3 are the most relevant. This can be explained by the fact that they are linear and quadratic in the complex moments as opposed to fourth order for the last invariant. Therefore they should capture the more general structure whereas A_4 should capture more fine details. Interestingly, among the

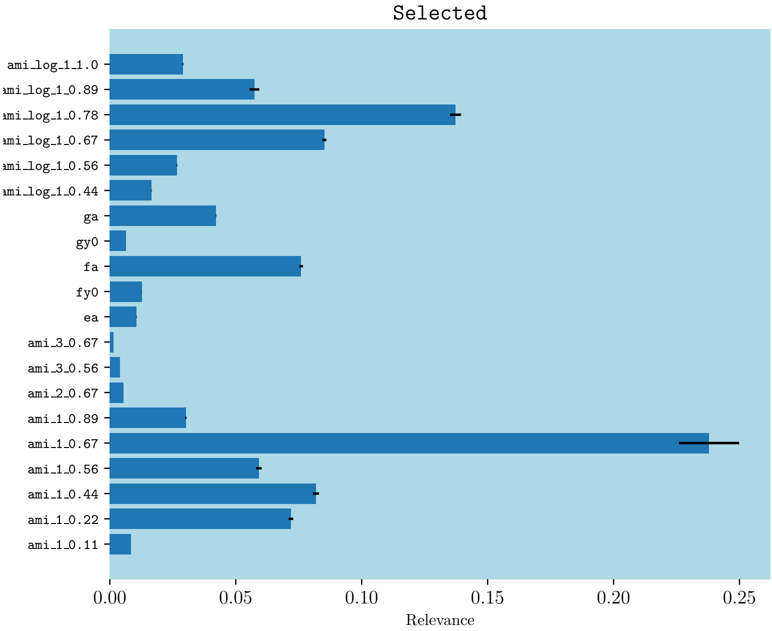
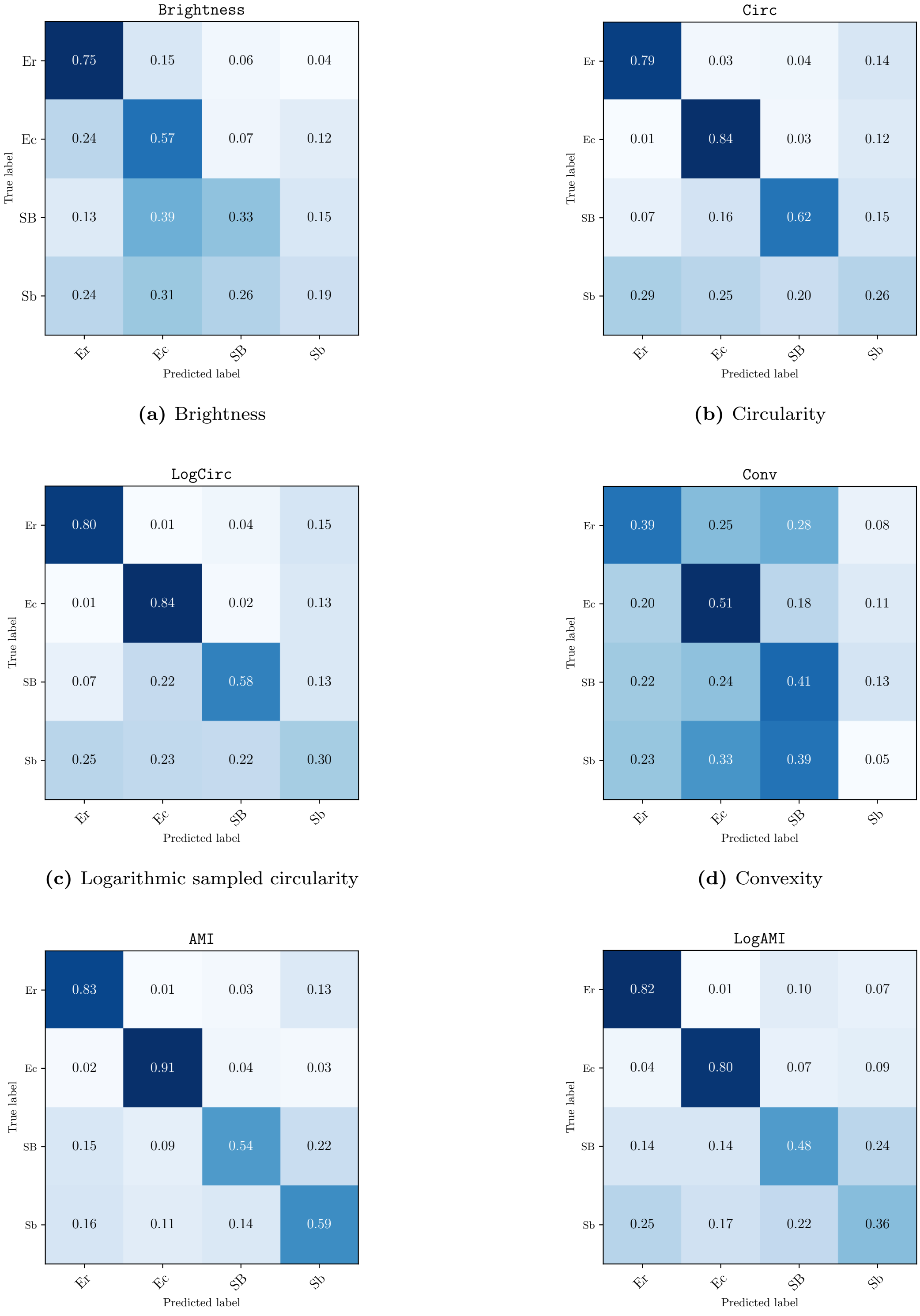


Figure 3.5: Feature relevances for the derived feature vector.

brightness features, the `ga, fb, eb` (the 'sersic' index) features, are found to be less relevant than the `ea, fa, ga, fy0, gy0` features.

Next, it was investigated if it was possible to improve on the AMI feature vector. This was done by iteratively including circularity, brightness, and logarithmically sampled features of the first AMI, running the classifier again and removing features with low relevance (less than 1% was chosen as threshold). This resulted in the derivation of the 'Selected' feature vector. It uses a subset features from the 1st, 2nd and 3d AMI, the `ea, fa, ga, fy0, gy0` features from brightness mentioned above and a subset of the logarithmically sampled 1st AMI for a total of 36 features. The full relevance profile with features names can be seen in figure 3.5. A score of 76.5% was reached using LGMLVQ using this feature vector. Note the high relevance for `fa` and `ami_1_0.67`. The circularity features did not improve the result at all. This is probably due to the fact that they are also derived from moments and therefore likely correlate with AMI features.



(a) Brightness

(b) Circularity

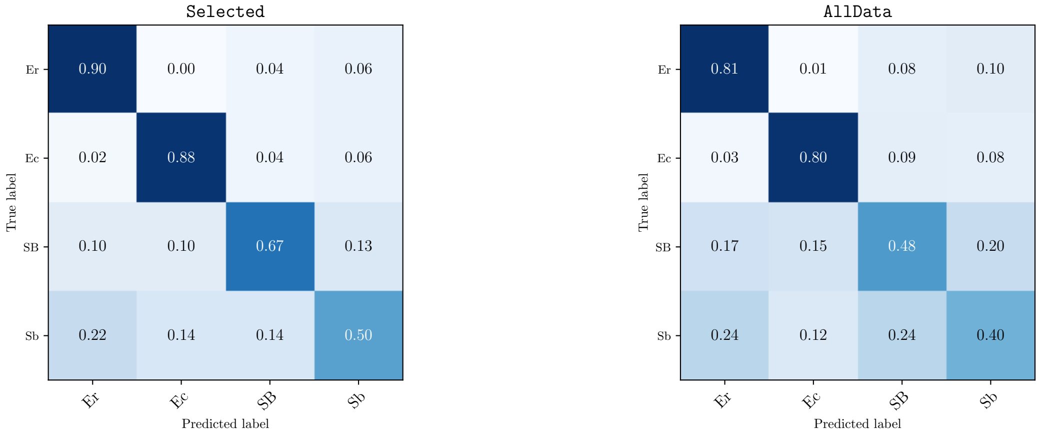
(c) Logarithmic sampled circularity

(d) Convexity

(e) AMI

(f) Logarithmic sampled AMI

Figure 3.6: Confusion matrices, normalized and averaged over 10-fold cross validation iterations. Matrix element $CM_{i,j}$ indicates the probability that the classifier predicts the label j where the true label is i.



(a) Derived feature vector

(b) Combined feature vector

Figure 3.7: Confusion matrices, contd.

Table 3.1: GMLVQ classifier scores in percentages.

Features	Dim.	Mean score (%)	Variance score (%)
Brightness	8	45.75	1.04
Circ	9	62	0.55
LogCirc	9	63	0.48
Conv	9	34	0.33
AMI	36	71.25	0.45
LogAMI	36	58	0.71
Selected	20	**74.5**	0.60
Combined	107	62.25	1.07

Table 3.2: LGMLVQ classifier scores in percentages.

Features	Dim.	Mean score (%)	Variance score (%)
Brightness	8	40.5	0.65
Circ	9	70.25	0.5
LogCirc	9	66.5	0.39
Conv	9	31.5	0.25
AMI	36	64.25	0.46
LogAMI	36	58	0.94
Selected	20	**76.5**	0.48
Combined	107	65.5	1.06

Chapter 4

Conclusions and future work

The use of a number of attribute profiles as feature vectors for galaxy classification has been tested. Experiments were done on a set of 400 galaxies classified by the GalaxyZoo2 project. The objects were automatically segmented using `MTObjects`, and moment and convexity based shape descriptors profiles were computed as feature vectors. In addition, the parameters of an analytic fit of the brightness profile was derived and evaluated as part of a feature vector. The viability of the features was explored and an optimized feature vector was derived. The best score of 76% was obtained using the LGMLVQ classifier. In particular, the results suggest that affine moment invariants and brightness profiles are suitable features for galaxy classification.

A simple but effective improvement to the feature vector could be made by taking more samples of the AMIs, and determining which of the sample points are the most significant.

Furthermore, the brightness profile fit discussed in section 3.4 can be improved by considering what the ideal points are at which to segment the profile, rather than always using the same points for every object; some objects with discontinuous brightness profiles were not able to be fit properly. The nature of the location of these points might also characterize useful information about the galaxy type. An alternative would be to also take samples of the brightness profile rather than fitting analytically.

Convexity features unexpectedly did not perform well during classification. As mentioned, the convexity profile is not computed in the same way as the other features and this could be improved upon, perhaps by augmenting the max-tree with information about its perimeter or exactly which pixels a node contains.

An ellipticity shape descriptor profile as a feature could be evaluated, such as the one proposed by Unić & Unić in [16]. Such a measure might be able to distinguish well between the basic classes of elliptical and spiral galaxies.

A lot of time is wasted during development because the program has to segment each FITS image before the GalaxyZoo objects are reached. This makes it time-consuming to try out new possible features or create new plots. This process could be sped up by precomputing a list of smaller `FITS` files which are centered on, and contain only a single given GalaxyZoo object. Since MTObjects relies on the presence of flat-tiles for background estimation, the computed background parameters should be saved for each FITS file, for example in the header. Other information such as an object's label could be stored as well.

Appendix A

Feature relevance profiles

In this appendix the learned relevance profiles are provided which reflect the diagonal of the relevance matrix of the GMLVQ method after training. The figures display the mean and variance of each relevance over 10-fold cross validation. It is important to note that relevance profiles are not unique, and a low relevance does not mean that a feature is unimportant. However, a relevance might be lower for a particular feature due to pairwise correlation among the features in the vector.

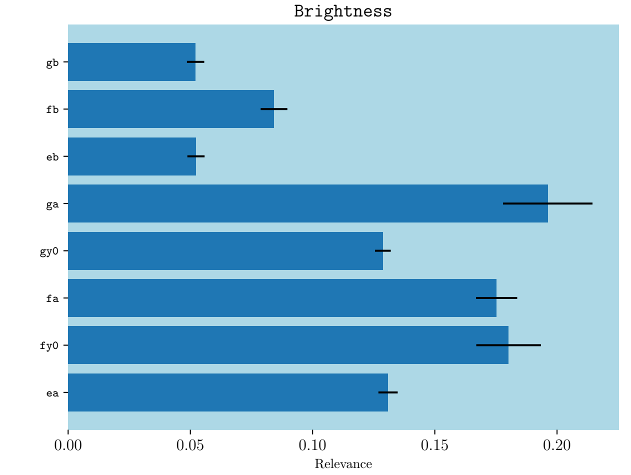


Figure A.1: Brightness feature vector relevance profile.

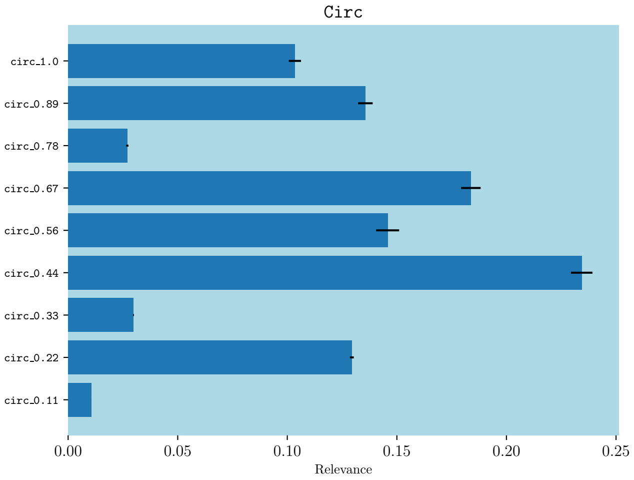


Figure A.2: Circularity feature vector relevance profile.

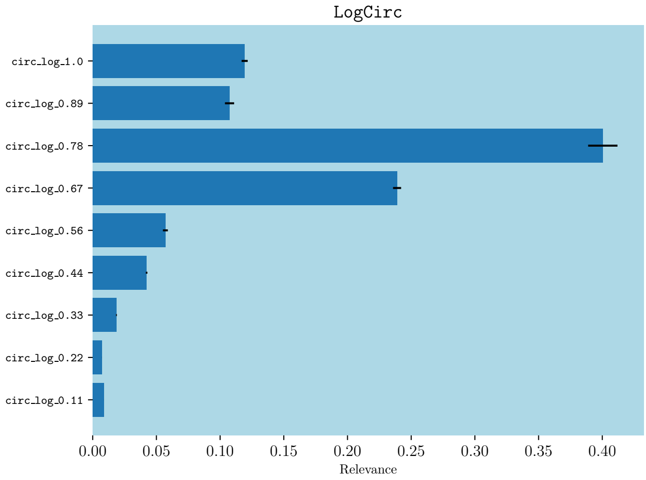


Figure A.3: Logarithmic scale circularity feature vector relevance profile.

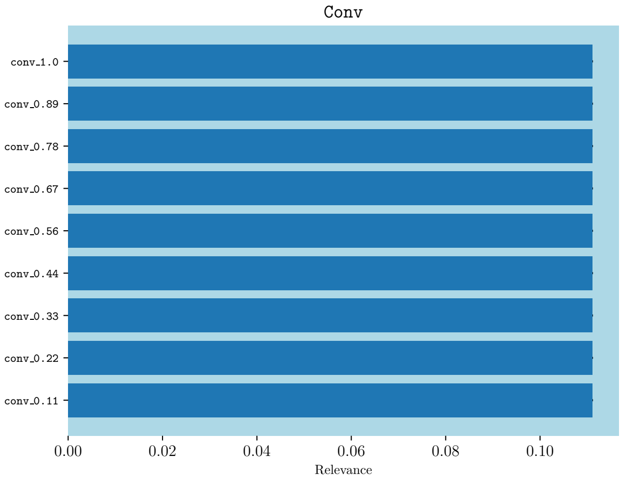


Figure A.4: Convexity feature vector relevance profile.

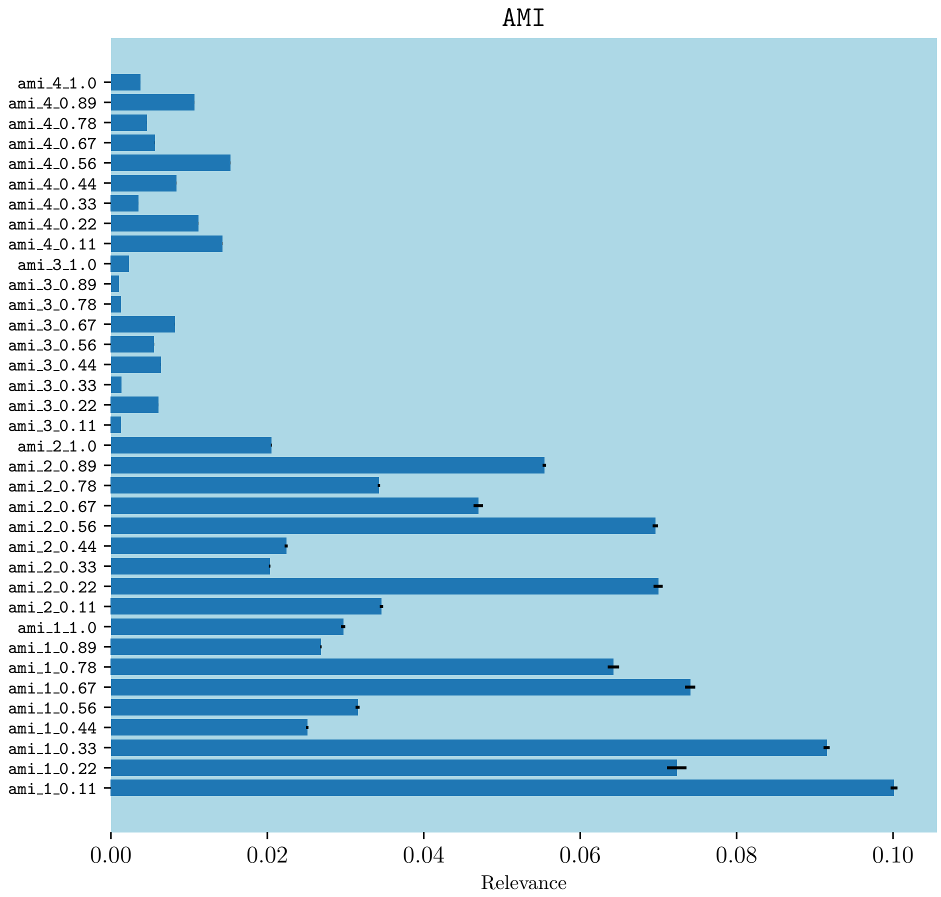


Figure A.5: AMI feature vector relevance profile.

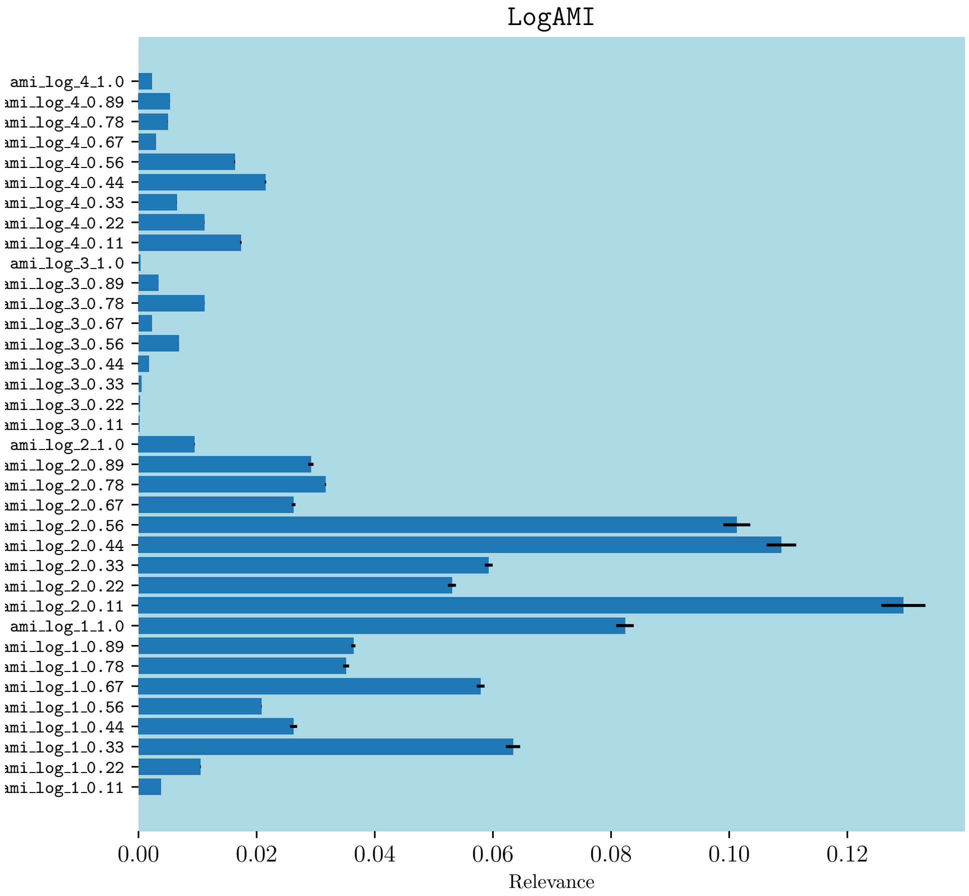


Figure A.6: Logarithmic scale AMI feature vector relevance profile.

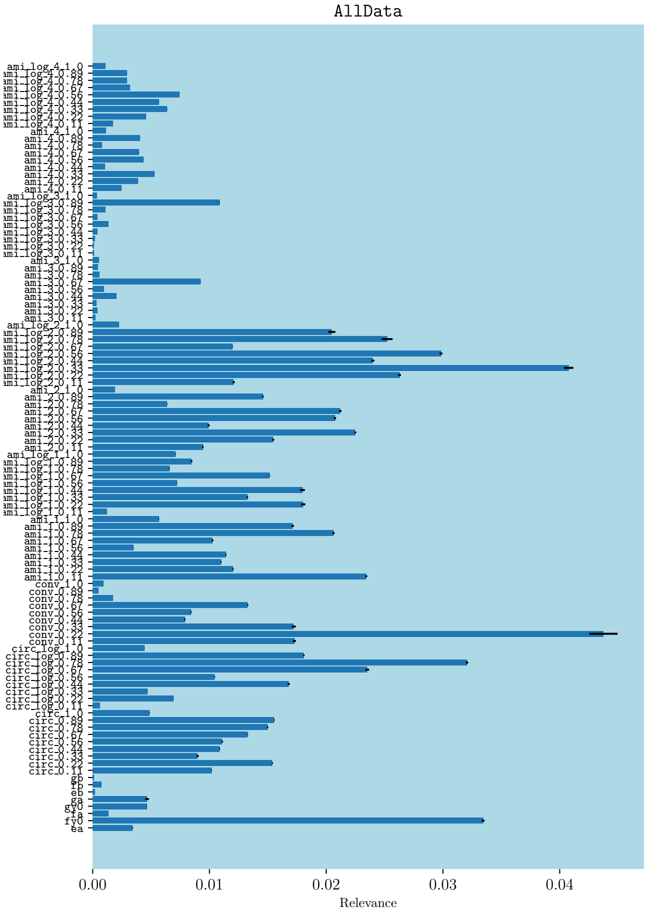


Figure A.7: Combined feature vector relevance profile.

Bibliography

[1] Hu, M.K.: *Visual pattern recognition by moment invariants.* IRE Transactions on Information Theory 8(2), 179–187 (1962)

[2] Flusser. J., Suk, T.: *The Independence of the Affine Moment Invariants.* AIP Conference Proceedings 860, 387 (2006)

[3] Nolte, A., Wang, L., Bilicki, M., Holwerda, B., Biehl, M.: *Galaxy classification: A machine learning analysis of GAMA catalogue data*, Elsevier (2019)

[4] Teeninga, P., Moschini, U., Trager, S.C., Wilkinson, M.H.F: *Statistical attribute Filtering to detect faint extended astronomical sources.* Math. Morphol. Theory Appl. 2016; 1:100–115 (2016)

[5] Kelvin, L.S. et. al: *Galaxy And Mass Assembly (GAMA): ugrizY JHK Sérsic luminosity functions and the cosmic spectral energy distribution by Hubble type.* Mon. Not. R. Astron. Soc. 000, 1–23 (2013)

[6] `https://data.galaxyzoo.org/`

[7] Hubble, E.P. : No. 324. *Extra-galactic nebulae.* Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington. 324: 1–49 (1926)

[8] A. Sato, K. Yamada. *Generalized Learning Vector Quantization.* Information Technology Research Laboratories (1996)

[9] De Vaucouleurs, G. *Classification and Morphology of External Galaxies.* Handbuch der Physik. 53: 275. (1959)

[10] *The Yerkes classification.* `https://ned.ipac.caltech.edu/level5/CLASSIFICATION/yc.html`

[11] https://github.com/CarolineHaigh/mtobjects

[12] P. Teeninga, U. Moschini, S. C. Trager, M.H.F. Wilkinson. *Statistical attribute filtering to detect faint extended astronomical sources* Math. Morphol. Theory Appl. 2016; 1:100-115

[13] J. L. Sérsic. *Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy.* (1963)

[14] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision.* Chapman and Hall (1993)

[15] J. Žunić, K. Hirota P. L.Rosin. *A Hu moment invariant as a shape circularity measure.* Pattern Recognition, Volume 43:47-57 (2010)

[16] D Unić, J. Unić *Shape ellipticity from Hu moment invariants.* Applied Mathematics and Computation, Volume 22:406-414 (2014)

[17] https://dr15.sdss.org/home

[18] Willett et al.*Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey.* Mon. Not. R. Astron. Soc. 000, 1–29 (2013)

[19] P. Salembier, A. Oliveras, L. Garrido, IEEE T. Image Process., 7, 555 (1998)

[20] U. Moschini, P. Teeninga. S. C. Trager, M.H.F. Wilkinson. *Parallel 2D Local Pattern Spectra of Invariant Moments for Galaxy Classification.* CAIP 2015: Computer Analysis of Images and Patterns pp 121-133 (2015)

[21] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, 1997.

[22] E. Rahtu. M. Salo. J. Heikkila. *A new convexity measure based on a probabilistic interpretation of images* IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume: 28 , Issue: 9 (Sept. 2006)

[23] P. Schneider, M. Biehl, B. Hammer. *Adaptive relevance matrices in learning vector quantization.* Neural Comput. 2009 Dec;21(12):3532-61 (2009)

[24] https://github.com/MrNuggelz/sklearn-lvq

[25] A. Meijster, M.H.F. Wilkinson. *A comparison of algorithms for connected set openings and closings.* Pattern Analysis and Machine Intelligence, IEEE Transactions on. 24. 484-494. (2002)