Automatic object detection and categorisation in deep astronomical imaging surveys using unsupervised machine learning

Author:
Alexander HOCKING

Supervised by:
Dr. Yi Sun PRIMARY & Dr. J. E. Geach PRIMARY
Dr. Neil Davey SECONDARY

Centre for Computer Science and Informatics Research
School of Computing
University of Hertfordshire

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

March 2018

Abstract

I present an unsupervised machine learning technique that automatically segments and labels galaxies in astronomical imaging surveys using only pixel data. Distinct from previous unsupervised machine learning approaches used in astronomy the technique uses no pre-selection or pre-filtering of target galaxy type to identify galaxies that are similar. I demonstrate the technique on the *Hubble Space Telescope (HST)* Frontier Fields. By training the algorithm using galaxies from one field (Abell 2744) and applying the result to another (MACS0416.1-2403), I show how the algorithm can cleanly separate early and late type galaxies without any form of pre-directed training for what an ‘early’ or ‘late’ type galaxy is. I present the results of testing the technique for generalisation and to identify its optimal configuration. I then apply the technique to the *HST* Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) fields, creating a catalogue of 60000 labelled galaxies, grouped by their similarity. I show how the automatically identified groups contain galaxies with similar morphological (and photometric) type. I compare the catalogue to human-classifications from the Galaxy Zoo: CANDELS project. Although there is not a direct mapping, I demonstrate a good level of concordance between them. I publicly release the catalogue and a corresponding visual catalogue and galaxy similarity search facility at www.galaxyml.uk. I show how the technique can be used to identify rarer objects and present lensed galaxy candidates from the CANDELS imaging. Finally, I consider how the technique can be improved and applied to future surveys to identify transient objects.

Declaration

I declare that no part of this work is being submitted concurrently for another award of the University or any other awarding body or institution. This thesis contains a substantial body of work that has not previously been submitted successfully for an award of the University or any other awarding body or institution.

The following parts of this submission have been published previously and/or undertaken as part of a previous degree or research programme:

1. Chapter 3: Sections 3.3.2, 3.3.3 and 3.4.2.1 were previously published in Hocking et al., 2018 *Monthly Notices of the Royal Astronomical Society*, 473, 1108
2. Chapter 4: Section 4.5 was previously published in Hocking et al., Unsupervised Image Analysis and Galaxy Categorisation in Multi-Wavelength Hubble Space Telescope Images *Proceedings of the European Conference on Machine Learning (ECML) Doctoral Consortium* 2015,
3. Chapter 5: The content of this chapter was previously published in Hocking et al., Mining Hubble Space Telescope Images, *Proceedings of the International Joint Conference on Neural Networks 2017 (IJCNN)*
4. Chapter 6: This content of this chapter was previously published in Hocking et al., 2018 *Monthly Notices of the Royal Astronomical Society*, 473, 1108

Except where indicated otherwise in the submission, the submission is my own work and has not previously been submitted successfully for any award.

Acknowledgements

I never intended to do a PhD, but I'm glad I did. The research involved is possibly the purest expression of innovation there is and that's what I call fun!

Before I started I wasn't sure what project to do other than 'something to do with machine learning'. I was always interested in Astronomy but it was not something I was remotely thinking of at the time. I met with Yi Sun and Neil Davey, and at about the same time they were approached by Jim Geach about a machine learning astronomy project. The course was set. It is true to say that supervisors and project choice can make or break a PhD. But, I honestly couldn't have been more fortunate here. The guidance from Jim, Yi and Neil has been brilliant. Thank you very much for the opportunity, your advice and your effort. I hope you guys enjoyed the journey as much as I did.

The camaraderie involved in research is immense. The fellow researchers you get to know along the way make the day to day experience of research really fun. I've made friends for life. So thank you to Ankur, Jean, Zaheed, Ed, Marco and Dimitris, it wouldn't have been the same without you guys.

My Mum & Dad didn't have the best scholastic experience! Dad left school at 15, and Mum at 16. It's just how it was for them - they were expected to go out to work. Study and qualifications, let alone University, was what other people did. It would have been very easy for them to project those same expectations on to their four children. But they did not. For most of my life I was completely unaware of the 'ruse' they played on me that doing well at school and going to University was normal. Mum & Dad, thank you for doing what you did.

There is, however, one person to blame for all the shenanigans over the past few years and that is my wife Nancy. I had been presented with a choice. I had successfully interviewed for a safe, well paying position, which was a fine but fairly unexciting job. The alternative was the pursuit of an MSc by Research - the opportunity to work in a professional research environment. An experience that was unlikely to be available in the future. Choosing to quit my existing job and reject the new one, felt, at the time, like jumping out of an aeroplane without a parachute. It felt like a crazy decision. But it was time to choose between what was meaningful and what was safe and expedient. At the moment of (in)decision my Wife provided me with a swift push in the right direction. Nancy, thank you for pushing me out of the aeroplane and for your unwavering support since.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Astronomers seek to understand the physical processes occurring within galaxies and how they evolve over cosmic time. A fundamental method astronomers apply to achieve this is the analysis of images. By studying images in different bands of the electromagnetic spectrum, astronomers can analyse galaxies of different types, ages and masses, how they interact and merge, their structure and composition and how nearby galaxies differ from distant galaxies. In this Chapter I will provide a brief review of existing techniques used in Astronomy related to my research. I shall identify my research question in Section 1.3.

1.1 An overview of image analysis in astronomy

1.1.1 Classification systems

In the 1920s, using the largest telescope of the time, the Hooker 100-inch reflector, Edwin Hubble classified galaxies into two main morphological types, spirals and ellipticals (Hubble, 1926). He also identified a third type called an irregular galaxy. Figure 1.1 shows individual examples of a spiral and elliptical galaxy.

Hubble developed this scheme further to produce what is known as the Hubble Tuning Fork or Hubble Sequence (Hubble, 1936). Figure 1.2 shows a modern version of the Hubble Sequence. The term “Tuning Fork” is a reference to the split of spiral galaxies into two groups: with and without a central bar. The sequence gives the impression of evolution, although there was no

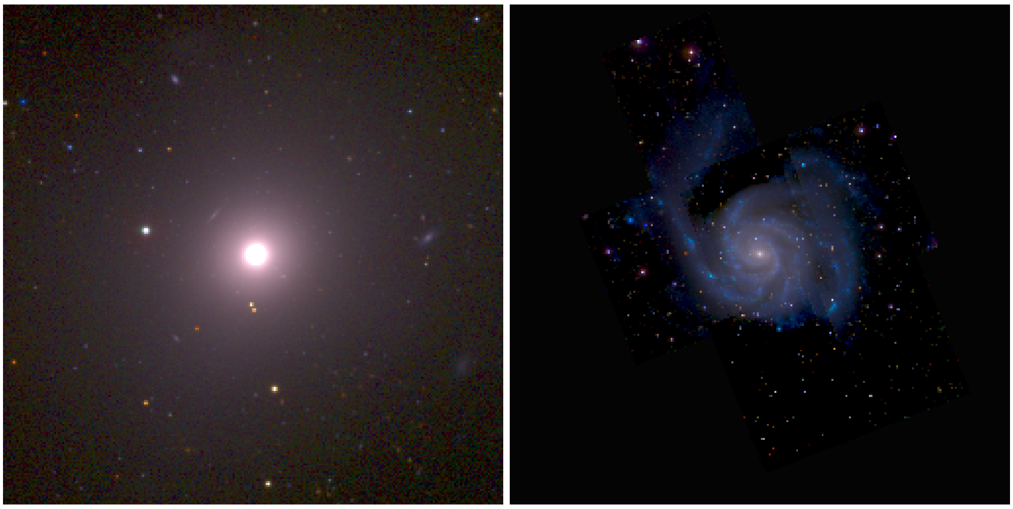


FIGURE 1.1: The elliptical galaxy M89 which is spheroid and spiral galaxy M101 known as the Pinwheel galaxy. Images taken by the Sloan Digital Sky Survey (SDSS) in bands r, g and i Baillard et al. (2011)

evidence to support this. Astronomers were simply able to classify galaxies in the local Universe using this system. Sandage (2005) describes the development of this classification scheme.

The original Hubble morphological scheme was developed using a series of photographic plates taken at the Mount Wilson Observatory (Hubble, 1926, 1936). De Vaucouleurs (1959, 1964) extended the original Hubble classification system to include further sub types, including a smooth transition beyond types Sc and SBc to the Irregulars. He introduced nomenclature to describe spirals without bars (SA) and intermediate types with weak bars (SAB), galaxies with rings (r), without rings (s) and intermediates (rs), as well as further classifications relating to the spiral arms, e.g. Sd (diffuse, broken arms with a faint central bulge), Im (highly irregular) and Sm (irregular with no bulge), effectively replacing Hubbles Irr classification. Holmberg (1958) also introduced plus and minus symbols to denote even finer divisions. In van den Bergh (1960a,b) added the concept of 'order' in the spiral arms, adding the labels I, I-II, II etc. to indicate increasing disorder within a Hubble type. These classifications also indicate increasing luminosity and the modern system covers a magnitude range of four between I and V (Sandage, 2005). More recently, detailed morphological features such as clumpiness and tidal tails are being classified by Galaxy Zoo (see Section 1.2.1).

More recent observations have imaged galaxies of higher cosmological redshift. Cosmological redshift arises as a consequence of photons of light travelling through an expanding Universe. The wavelength of the photons expands in direct proportion to the expansion of the Universe they travel through resulting in redshift z defined in equation 1.1, where λ is wavelength.

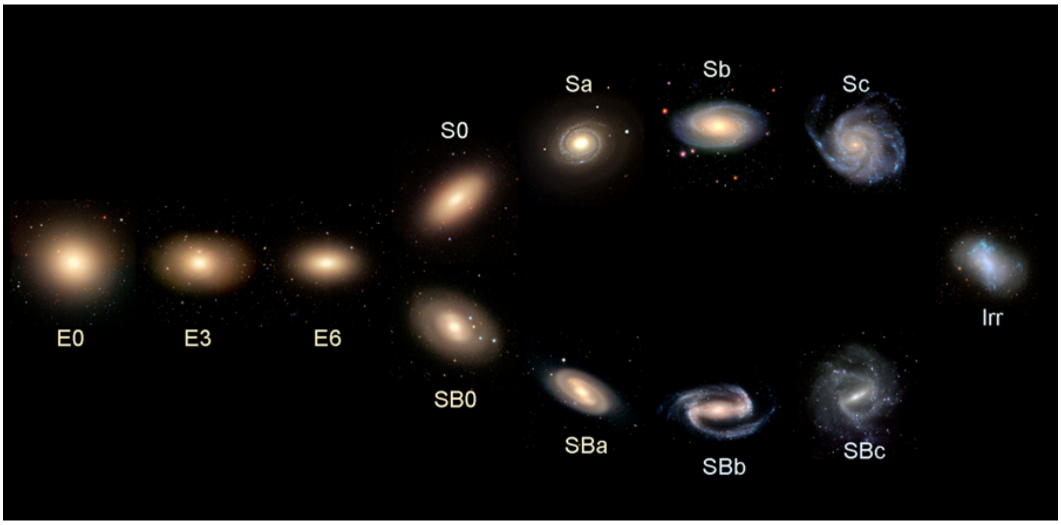


FIGURE 1.2: The Hubble Tuning Fork morphological classification scheme. Elliptical galaxies are to the left. They are numbered '0' to '6' with '0' being spherical and '6' being very elliptical. The spiral galaxies are to the right. The letters 'a', 'b' and 'c' signify a combination of the central bulge size and the compactness of the spiral arms. The spiral galaxies are ordered into two distinct types, the top fork shows ordinary spiral galaxies, and the bottom fork shows spiral galaxies that have a central bar. The 'S0' galaxies are known as lenticular galaxies. The subtypes were added to the classification system by De Vaucouleurs (1959). Irregular galaxies 'Irr' have irregular morphologies and are difficult to place in the diagram, but were positioned to the right of the sequence. Credit: Mortlock (2013)

$$z+1=\frac{\lambda_{observed}}{\lambda_{emitted}} \tag{1.1}$$

We cannot directly measure distance nor age directly, and therefore we use the measurable cosmological redshift as a proxy (Shu, 1982). Galaxies of high redshift appear as they did at a remote time when the Universe was much smaller than it is now. These galaxies are typically at an earlier stage of evolution and recent observations indicate that they cannot be classified as a single Hubble type and a new classification scheme may be required (Mortlock et al., 2013; Conselice, 2014).

1.1.2 Quantitative measures

1.1.2.1 Quantifying the light-profile of galaxies

The Hubble morphological classifications of galaxies are primarily a descriptive taxonomy distinguishing particular galaxy types. Quantitive measures to analyse, categorise and classify individual galaxies in images have also developed. De Vaucouleurs' law was one of the first models of how the surface brightness of an elliptical galaxy varies as a function of distance

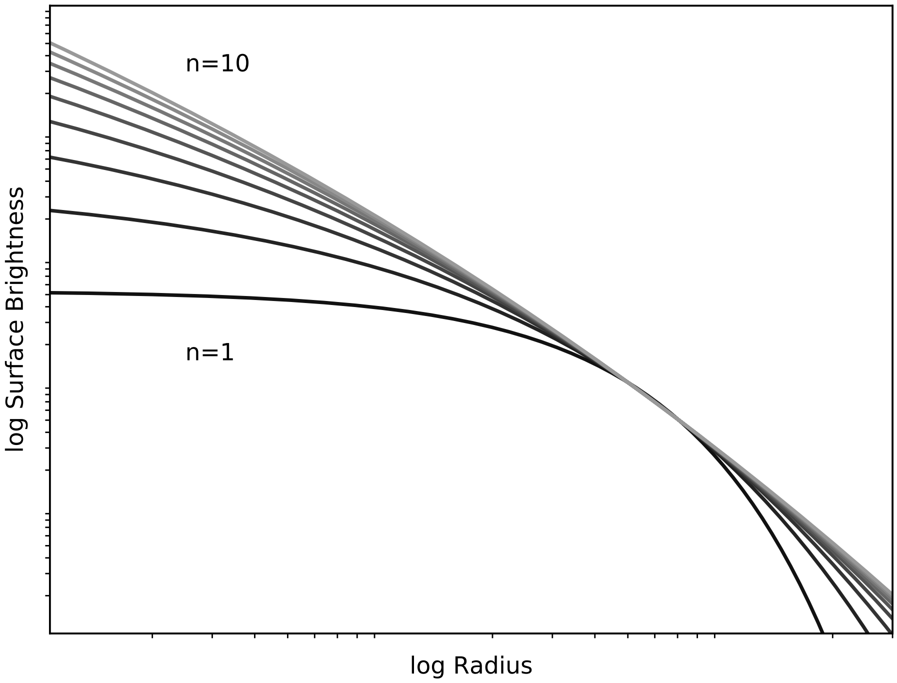


FIGURE 1.3: Ten Sérsic 1D profiles each with a different Sérsic index, n.

from its centre (de Vaucouleurs, 1948). De Vaucouleurs also noted that the disk component of many galaxies could be described by an exponential model (de Vaucouleurs, 1948; Freeman, 1970). Sérsic developed a more general form of the De Vaucouleurs' law called the Sérsic profile (Sérsic, 1968). This is described in equation (1.2) where $I(r)$ is the intensity of a galaxy as a function of projected radius r from its centre, I_0 is the light intensity at $r = 0$, α is the scale length corresponding to the radius where the intensity drops by $\frac{1}{e}$. The parameter n is known as the Sérsic index and describes the profile shape, for example, using $n = 4$ reproduces de Vaucouleurs' law and $n = 1$ reproduces the exponential model. Figure 1.3 shows the Sérsic profiles for $n = 1$ to $n = 10$.

$$I(r) = I_0 \exp\left(-\left(\frac{r}{\alpha}\right)^{1/n}\right) \tag{1.2}$$

These parametric models have been used extensively to perform structural studies of the 1d and 2d light profiles of galaxies. There has been a significant amount of work fitting Sérsic profiles to the profiles of nearby galaxies (Caon et al., 1993; Kormendy et al., 2009; Graham, 2013). A common approach is to model a galaxy's central bulge and disk separately, known as bulge/disk decomposition (Kormendy, 1977; Caon et al., 1993). Figure 1.4 shows an example of a bulge/disk decomposition. Three model profiles are fitted, representing the disk, central bulge and bar. Subtracting these from the original image creates a 'residual' which reveals how well the functions model these components of the galaxy and also highlight features such as dust lanes. More recently it was discovered that galaxy components may not have the same centre, for example, the existence of offset disks and bars (Kruk et al., 2017).

This approach has limitations due to assumptions such as that galaxies only have a single centre, and the ellipticity and angle of each component do not change with increasing distance from the galactic centre (Conselice, 2014). Also, many galaxies exhibit more than two components and so the decision of which galaxy components to fit requires visual inspection. The extensive morphological catalogues produced by Galaxy Zoo can be used here to automatically choose which components to fit (Willett et al., 2013, 2016; Simmons et al., 2016a). However, evaluating the residual image is a manual effort. Fitting the light profile of galaxies and their components is easier for small samples of nearby galaxies where it is possible to use an interactive fitting process. Tools have been developed to automate this process further (Häußler et al., 2013). Kruk et al. (2018) successfully combined Galaxy Zoo morphological data and GALFITM, enhanced as part of the MegaMorph project (Häußler et al., 2013), to study secular evolution of barred galaxies.

1.1.2.2 Quantifying the structure of galaxies

Perhaps the most popular methods of quantifying structure are the combination of concentration (C), asymmetry (A) and clumpiness (S) commonly known as the CAS system (Conselice, 2003); the Gini, and M_{20} measures (Lotz et al., 2004) and more recent measures by Freeman et al. (2013). These capture major features of galaxies, such as symmetry and clumpiness, without the need to make the sort of assumptions required by parametric light profile fitting methods such as the Sérsic profile. I briefly review the CAS parameters to provide an idea of the concept behind these techniques. I give a brief overview of Conselice (2014) and Figure 1.5 provides a visual illustration of applying CAS to a galaxy:

- Concentration (C) – a measure of the radial distribution of flux. One simple method for calculating C, defined by Kent (1985) is to use the ratio of two radii containing, for

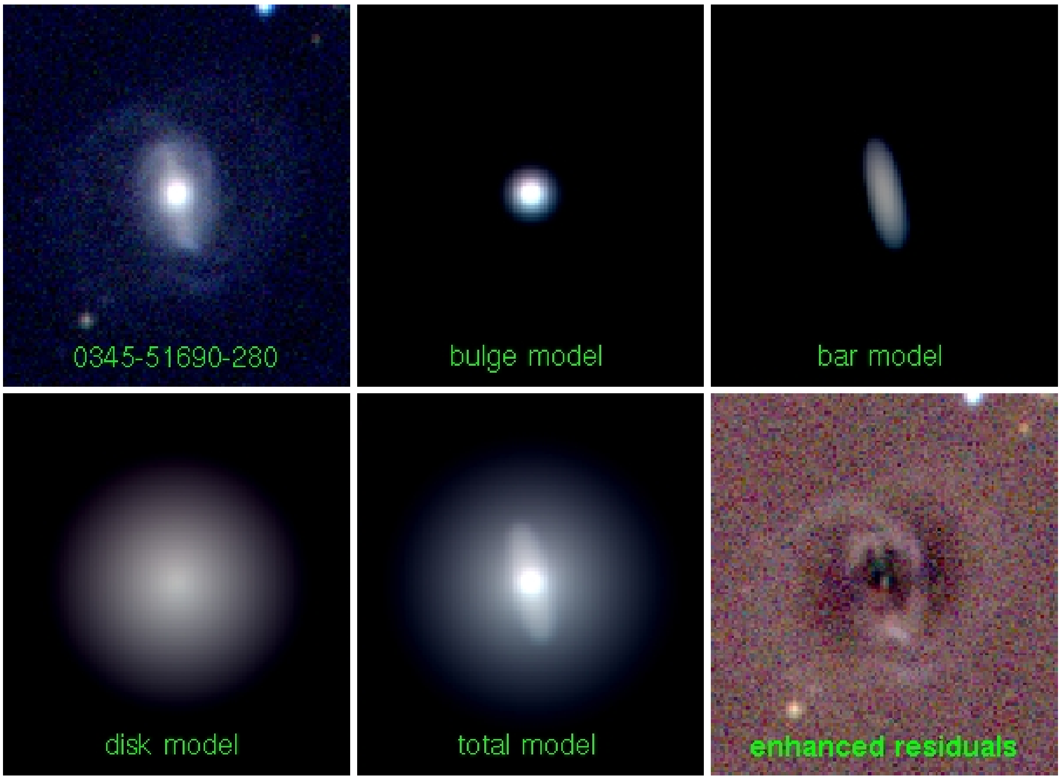


FIGURE 1.4: The modelling of the light profile of a galaxy and its components. The top left image is the original grey scale image of a galaxy. Three galaxy components were fitted: the top central image is the model of the central bulge, the top right image is the model for the central bar, the bottom left image is the model of the disk, the bottom middle image is the total model image which is a combination of the bulge, disk and bar models. The bottom right image shows the enhanced residual after the total model is subtracted from the original image. Credit: ESO/E. De Souza and D. Gadotti/BUlge Disk Decomposition Analysis tool (BUDDA).

example, 20% (r_{inner}) and 80% (r_{outer}) of the total galaxy flux.

$$C = 5 \times \log_{10}\left(\frac{r_{\text{outer}}}{r_{\text{inner}}}\right) \tag{1.3}$$

- Asymmetry (A) – A galaxy image I_0 is rotated 180 degrees around its centre I_{180} and the pixel values are then subtracted from the original image. I_0 represents the pixel intensities of the original galaxy image and I_{180} represents the pixel intensities of the image after rotating it 180 degrees, B_0 is a blank area of sky near the galaxy. The *min* refers to the global minimum found in an iterative process required to identify the centre of rotation. An initial guess is made and the left and right parts of the equation are calculated for this centre and for the eight surrounding points. The points that produce the minimum values are used to calculate the final value for A. Asymmetry is effective for distinguishing types of galaxies, for example, elliptical galaxies, being very symmetric have $A \sim 0.02 \pm 0.02$ and late-type spiral (Sc-Sd) galaxies $A \sim 0.17 \pm 0.1$ (Conselice, 2014). The following equation is given in Conselice (2014). The summations are applied to all of the pixel intensities in each 2D image matrix: $I_0, I_{180}, B_0, B_{180}$.

$$A = min\left(\frac{\sum |I_0 - I_{180}|}{\sum |I_0|}\right) - min\left(\frac{\sum |B_0 - B_{180}|}{\sum |I_0|}\right) \tag{1.4}$$

- Clumpiness (S) – is calculated by subtracting a smoothed version of a galaxy from the original image. It is a measure of smoothness. Elliptical galaxies are smooth whereas star-forming regions in galaxies typically appear very clumpy. $I_{x,y}$ is the original image, $B_{x,y}$ is the smoothed image and the σ smoothing kernel. The following is given in (Conselice, 2014). The summations apply to the pixel intensities within the $I_{x,y}, I^{\sigma}_{x,y}, B_{x,y}, B^{\sigma}_{x,y}$ images.

$$S = 10 \times \left[\left(\frac{\sum(I_{x,y} - I^{\sigma}_{x,y})}{\sum I_{x,y}}\right) - \left(\frac{\sum(B_{x,y} - B^{\sigma}_{x,y})}{\sum I_{x,y}}\right)\right] \tag{1.5}$$

Of great concern for any of these techniques is the ability to compare measurements when applying them to galaxies at different redshifts. To ensure comparable results an effective and common definition for the pixels that form a galaxy is required in any measurement. The most common choice when using CAS is the Petrosian radius (Petrosian, 1976; Conselice, 2014). The combination of CAS parameters and the Petrosian radius has been found to be effective when comparing over broad redshift ranges, whereas Sérsic profiles and other radius measures use assumptions or are affected by measurement effects such as surface brightness dimming, which render comparisons across redshift ranges very difficult Conselice (2014). The Petrosian radius, defined by the equation (1.6), is identified when the intensity at the radius r_p is equal to the mean surface brightness within r_p multiplied by a threshold value η.

$$\mu(r_p) = \eta \times \overline{\mu}(r < r_p) \tag{1.6}$$

Two further measures, in addition to CAS, are used to quantify galaxies are the Gini coefficient (Gini, 1912) and M_{20} (Lotz et al., 2004). These measures are sometimes used in combination with the CAS parameters, for example, in Lotz et al. (2008).

- The Gini coefficient G, as applied to astronomy (Abraham et al., 2003; Lotz et al., 2008), is based on the ordered cumulative distribution function of a galaxy's pixel values. Defined in Equation (1.7), where X_i is the pixel values, $\overline{X}$ is the mean pixel value, n is the

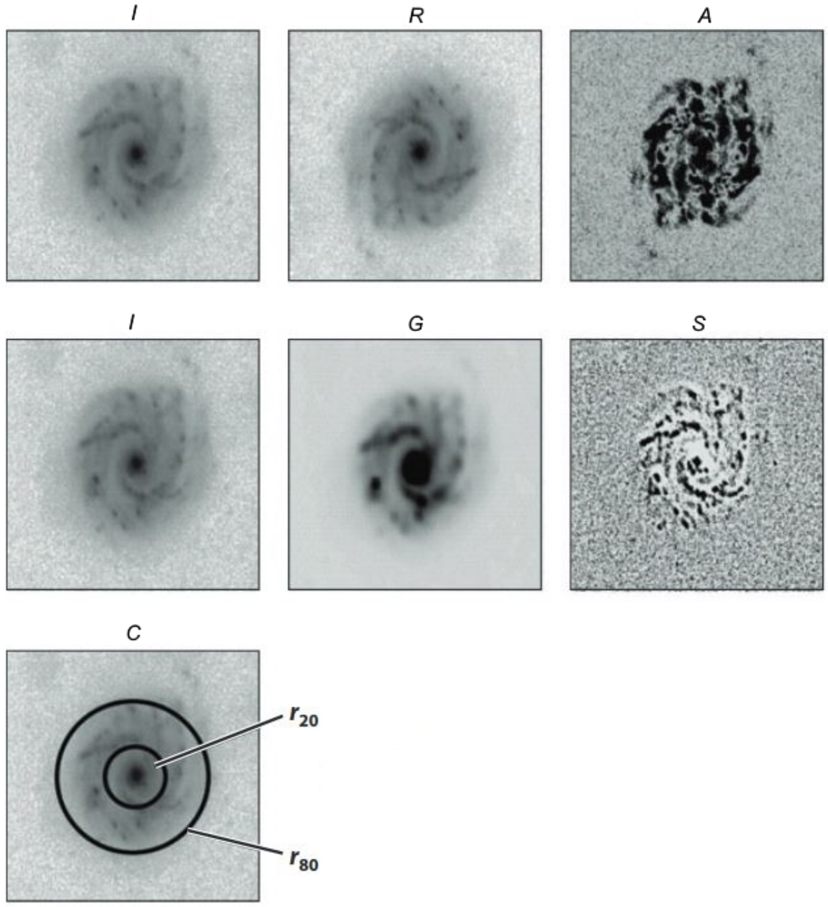


FIGURE 1.5: An example of the application of Concentration (C), Asymmetry (A) and Clumpiness (S) to measure the structure of a galaxy. The images to the left are the original galaxy. The middle column contains the 180 degree rotation (top) and the blurred version of the galaxy (middle), the two right-hand images are the residual images. The bottom image is shows example radii used to calculate Concentration. Credit: Based on *Conselice (2014)*

number of pixels in the galaxy image (defined by the Petrosian radius). This coefficient represents the mean of the absolute difference between all pairwise combinations of X_i. A coefficient value of 0 indicates a uniform galaxy profile and a value of 1 all the light is located in a single pixel.

$$G = \frac{1}{2\overline{X}n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j| \tag{1.7}$$

- M_{20} is defined as the normalized second order moment of the brightest 20% of the galaxy's flux (W/m^2), see equation (1.9). The Petrosian radius is typically used to define which pixels belong to the galaxy in the image. M_{20} is calculated by ordering the galaxy's pixels by flux, sum M_i over the brightest pixels until the sum is equal to 20% of the total flux and then normalise with M_{tot} to remove the dependency on total galaxy flux (or galaxy size). The total second-order moment M_{tot} (see equation 1.8) is the flux in each pixel f_i multiplied by the squared distance to the centre of the galaxy, summed over all the galaxy pixels. Where f_i is the flux of a pixel i and x_c, y_c is the galaxy's centre. The centre is found by minimizing M_{tot}.

$$M_{tot} = \sum_{i=1}^{n} M_i = \sum_{i=1}^{n} f_i \cdot ((x_i - x_c)^2 + (y_i - y_c)^2) \tag{1.8}$$

$$M_{20} = \log 10 \left(\frac{\sum_i M_i}{M_{tot}} \right) \text{while} \sum_{i=1} f_i < 0.2 f_{tot} \tag{1.9}$$

1.1.3 Combining features to identify properties and morphological types

Conselice (2003) discovered that the combination of CAS parameters formed a three dimensional parameter space that could be used to identify different morphological types of galaxy. For example, galaxies with a high light concentration, low asymmetry and low clumpiness are likely to be elliptical galaxies. Combinations of CAS values that could identify several other morphological types were established (Conselice, 2003).

Another useful combination of parameters was identified by Abraham et al. (2003). The Gini coefficient (Lotz et al., 2004), central concentration, and mean surface brightness, when sampled from nearby galaxies, were found to form a 3D-plane in parameter space.

Another example is the combination of parameters to form what astronomers call the fundamental plane of elliptical galaxies (Sandage, 2005). The fundamental plane is a relationship between the effective radius (the radius that contains half the total light of the galaxy), average surface brightness and central velocity dispersion (the σ of the radial velocities of the stellar population in the interior of a galaxy) of normal elliptical galaxies. Any one of the three parameters may be estimated from the other two, as together they describe a plane that falls within the three dimensional parameter space.

All of these measures, with the exception of velocity dispersion, can be applied to images of galaxies to provide a quantitative measure of galaxies properties. They can be combined to form parameter spaces. Using the terminology of machine learning, these measures are an individual 'feature'. The Fundamental Plane has three features, and CAS has three features. Together these features are combined to form a data manifold.

Although CAS, Gini and M20 have been popular, there are deficiencies. In particular, (Conselice, 2003, 2014) published ranges of values for each of the CAS parameters for several morphological types. These values were measured in using specific data sets. However, the presented average values have large uncertainties leading to considerable overlap across morphological types. For example, the uncertainties for Irregular galaxies, edge-on disks and Ultra Luminous Infra-red Galaxies (ULIRGs) overlap significantly. It is also not clear why these values would remain consistent across observations from different telescopes or images of significantly different depths from the same telescope. The classifications are also limited to nearby objects.

Also, each of the CAS parameters reduces an individual galaxy to a single value based on a broad measure across the whole of the galaxy. Therefore, no consideration is made for different components within galaxies. The limitations of this method are perhaps not surprising considering they were introduced in 2003, when computational resources were limited. It is now possible to analyse and create much more detailed models of galaxies. We are no longer limited to reducing an image of a galaxy to an ‘encoding’ of a few values such as the three CAS values. Instead, we can employ more complicated machine learning techniques such as Convolutional Neural Networks (see Section 1.2.3.1) and the model developed in this thesis. These machine learning models can encode the pixels of a galaxy into many more parameters allowing much more detailed information about galaxies to be retained and compared.

1.1.4 Surveys

There are two main types of imaging survey: images of a large region of sky, and images of several areas of sky that contain a known type of object. In this thesis the data I used considers both types.

Initial surveys performed in the early to mid 20th century were very limited when compared to the modern surveys of the last 30 years and those currently in development. Typically the numbers of galaxies used in research numbered in the hundreds or thousands (Djorgovski et al., 2013). There are now many more surveys and techniques that have been developed to image larger and larger areas of the sky. Perhaps the most famous large survey in visible light is the Sloan Digital Sky Survey (e.g. York et al., 2000; Stoughton et al., 2002; Abazajian et al., 2009; Blanton et al., 2017, SDSS). Using a 2.5-metre wide-field telescope the survey imaged over

10,000 deg^2 of the sky[1]. Its primary goal was to obtain images in five broad optical bands and obtain spectroscopy of over a million galaxies. The survey took ten years and the final data release (Abazajian et al., 2009), delivered in 2009, included almost a million galaxies and half a million stars.

Today, the SDSS is considered a large survey. But telescope and camera technology is improving very quickly. For example, in 2020, the space based Euclid telescope (Laureijs et al., 2011) will be launched. It is designed to observe 10 billion objects using optical (550 (green) to 920nm) and near-infrared (1000-2000nm) cameras.

Large, new ground-based telescopes are also arriving soon. The Large Synoptic Sky Telescope (Ivezic et al., 2014, LSST) will deliver perhaps the most ambitious optical survey of all. It will use a 3.2 gigapixel CCD camera to image 10,000 square degrees of the sky every three nights. The considerable survey area of 18,000 deg^2 will be imaged over 800 times in six optical bands (320nm to 1050nm). The telescope will image 20 billion galaxies and 20 billion stars, significantly more than previous surveys. One of the goals of the project is to 'make a high-definition colour movie of the deep Universe' (Ivezic et al., 2014). One of the most significant technical risks for the project is the availability of automated data analysis tools capable of processing the data quickly and accurately enough.

Although the LSST survey and Euclid will image billions of galaxies the vast majority of these galaxies will be too small or too dim for the telescopes to resolve morphological components. Therefore, morphological classification will only be possible for a subset of the total galaxies identified by these surveys.

1.2 Automated methods for analysing surveys

1.2.1 Galaxy Zoo: Crowd-sourcing visual classifications

The initial stages of most research into galaxy morphology has involved professional astronomers classifying hundreds to low thousands of galaxies. One of the largest manual efforts is a team of 65 astronomers who classified galaxies in the GOODS-South field of the Cosmic Assembly Near-infrared Dark Energy Legacy Survey (Kartaltepe et al., 2015, CANDELS).

[1] For example, 100 degrees × 100 degrees

To enable the visual classification process to be applied to much bigger datasets many more people are needed. This has led to the development of crowd-sourcing techniques and the Galaxy Zoo and Zooniverse projects (Lintott et al., 2008, 2010, GZ). GZ enables citizen scientists to participate in classifying surveys by using a website to classify galaxies sourced from surveys such as the SDSS and CANDELS. The citizen scientists, with typically limited knowledge of astronomy, are guided through the classification process using a decision tree of questions. Importantly, classification errors can be quantified as many classifications are obtained for each galaxy. GZ continues to classify huge numbers of galaxies from multiple surveys (Willett et al., 2013; Simmons et al., 2016a; Willett et al., 2016).

A key concern is ensuring consistent classification over time by individual citizen scientists and the weeding-out or down-weighting of classifications by unreliable classifiers (Simmons et al., 2016a).

A current development of the GZ and Zooniverse projects involves employing supervised machine learning techniques (see Section 1.2.3) in combination with citizen scientist classifications. An example of development in this direction is the combination of classifications from the Zooniverse Supernova Hunters project and a supervised machine learning algorithm to help identify supernovae in imaging from Pan-STARRS Survey for Transients Wright et al. (2017). Further development of these ideas focus on the optimum combination of citizen scientist and machine learning system in terms of classification accuracy and speed of classification (Beck et al., 2018).

1.2.2 Quantitative measures

Researchers have continued to develop tools and techniques to automate the analysis of surveys. In particular, tools to automate the calculation of structural parameters such as light profile modelling (see Figure 1.4) and non-parametric modelling such as CAS and Gini/M_{20} (see Section 1.1.2.2 and equations 1.3, 1.7 and 1.9).

Prominent tools for profile fitting are GIM2D (Simard, 1998) and GALFIT (Peng et al., 2002). These tools perform automatic quantitative morphology analysis by decomposing all objects in an input image. An example of this process is seen in Figure 1.4 in Section 1.1.2.1. The method has been successfully applied to perform bulge/disk decompositions of the galaxies in the SDSS survey (Simard et al., 2011). The user specifies up front which functions (such as a

Sérsic profile, or an exponential) to fit to the light profile. Estimating the number of components is a trial and error process using chi-squared and visually inspecting the pattern of residuals to identify the best fit. However, for fitting simpler models such as for bulge disc decomposition, there is little or no human interaction allowing their use for automated survey analysis.

The Galapagos tool (Barden et al., 2012) is a data pipeline that automates the application of GALFIT for simpler decompositions. Van der Wel et al. (2012) used Galapagos to perform structural decompositions of the CANDELS survey (See Section 2.5).

Other examples are Megamorph (Häußler et al., 2013), a project to extend GALFIT to fit models to multi-wavelength data. Also, PyMorph (Vikram et al., 2010) a tool to automate object location (source extraction), model fitting with GALFIT and the calculation of CAS and Gini/M_{20}.

Some tools employ image analysis techniques to identify and categorise individual galaxy components. Ganalyzer studies spiral structure (Shamir, 2011). This tool computes the slopes of the peaks detected in radial intensity plots to measure the spirality of a galaxy and determine its morphological class. The authors claim that this is a difficult manual task and in many cases the tool provides a more accurate analysis.

SpArcFiRe (Davis and Hayes, 2014) automatically identifies and categorises spiral arm segments. It categorises arm segments by using a least squares fit to a logarithmic spiral arc. It has been run on over 600,000 galaxies in the SDSS that are larger than 40 pixels across. They found a very good correlation between the quantitative description of the spiral structure and Galaxy Zoo classifications. Hart et al. (2017) used SpArcFiRe in combination with Galaxy Zoo data to constrain the mechanisms of spiral arm formation.

Automated tools, such as Galapagos, now resemble full data pipelines that contain source extraction, masking, segmentation, and object categorisation.

The advantage of these tools is that they allow the fast analysis of a large dataset. Though, whether the automation of these tools provides more accurate results is questionable.

The automation of structural decomposition tools has proved to be difficult. The main issue is that choosing which structural models to fit is very difficult without apriori knowledge of which structural components a galaxy has. For example, Allen et al. (2006) performed bulge and disk decomposition of over 10,000 galaxies and found that although many bulge and disc model fits were successful, they also ’automatically generated a lot of rubbish’. Simard et al. (2011) applied automation to analyse over a million SDSS galaxies. They restricted the analysis to a simple bulge and disk model fit and explained that using a more complicated model is not possible ’without compromising convergence and avoiding parameter degeneracies’.

Therefore, to ensure a good model fit, human inspection is still required in order to know which galaxy subcomponents are present before deciding which models to fit. Of great help here is the use of Galaxy Zoo catalogue data as it contains this structural information. For example (Kruk et al., 2017) looked for evidence of secular evolution in barred spirals by first using Galaxy Zoo data to identify a sample of SDSS galaxies with strong bars. However, even with this higher quality sample of galaxies, it was still necessary to inspect the galaxy image, model and residual to identify which model fit was relevant. Of the initial sample of 5282 barred galaxies, 3466 had meaningful model fits.

More flexible and accurate models that can automatically analyse more complicated galaxy morphology are needed. Machine learning applied to image data holds much promise for developing models with these characteristics.

1.2.3 Machine learning in astronomy

Machine learning started to appear as a bonafide field of academic research in the 1950s. A 1959 paper by Arthur Samuel of IBM (”Some studies in machine learning using the game of checkers”) is possibly one of the first pieces of published research to use this term (Samuel, 2000).

Today machine learning algorithms and techniques are becoming increasingly important for the efficient analysis of current and future astronomical surveys. As described in Section 1.1.4 future surveys will generate more data than is practical for humans to examine exhaustively. Moreover, existing techniques are limited in their capacity to represent galaxies when applied in an automated fashion, for example, detailed structural decomposition is still a manual effort (see Section 1.2.2).

For experiments such as the Large Synoptic Survey Telescope (Ivezic et al., 2014, LSST), it will be important to rapidly and automatically analyse streams of imaging data to identify interesting transient phenomena and to mine the imaging data for rare sources which may yield discoveries.

Machine learning has three broad categories: supervised, unsupervised and reinforcement learning (Bishop, 2006). I will review the applications of supervised and unsupervised learning to astronomy.

Reinforcement learning (Sutton and Barto, 1998) is currently under-represented in astronomy. However, its use has gained much recent publicity for automatically learning to play simple Atari games and to beat the world's best Go player (Mnih et al., 2015; Silver et al., 2016). The basic idea behind reinforcement learning is an agent will continually perform some form of action or actions and win a reward for performing the actions successfully. By repeatedly performing the actions typically many hundreds of thousands or millions of times the agent learns the parameter space of behaviours that maximise the reward. The analysis of astronomy survey data does not have an interactive component and it is not clear what type of reward could be defined. These are the likely reasons that reinforcement learning has not yet been applied to astronomy.

Supervised and Unsupervised machine learning techniques are used extensively in astronomy. The main difference between these two types of algorithms is that supervised machine learning algorithms require a training dataset with a set of outcomes or target measurements. The training dataset is used to build a prediction model that can predict the target value for new unseen data (Hastie et al., 2009). Unsupervised machine learning uses data without labels. Instead, the task is to reveal the underlying structure of data and how it is organised. By modelling the underlying structure it is possible to analyse new data for similar structure. I now consider the application of supervised and unsupervised algorithms in astronomy.

1.2.3.1 Supervised techniques

Supervised machine learning is used successfully in astronomy for both mundane and complex tasks. For example, there has been a significant effort made to develop techniques to improve the estimation of photometric redshifts. A photometric redshift is an estimate of cosmological redshift z (see Section 1.1.1). The estimate is calculated by measuring the brightness of an object at different wavelengths and comparing to model galaxy templates. There are several examples of neural networks being applied to this problem using labelled datasets (Firth et al., 2003; Collister and Lahav, 2004; Ball et al., 2004; Cavuoti et al., 2012; Brescia et al., 2013).

Bonfield et al. (2010) used Gaussian process regression (GP) to estimate redshifts and compared the results for small training sets. They found GP to be superior to the neural network approaches used in the ANNz tool (Collister and Lahav, 2004). More recent results, in the field of machine learning, reveal that neural networks are sensitive to training set size and typically have much higher accuracy when trained with larger datasets (LeCun et al., 2015).

A variety of supervised machine learning algorithms have been used to automatically classify objects such as stars and galaxies of different types, for example, by using neural networks (Klusch and Napiwotzki, 1993; Nielsen and Odewahn, 1994; Lahav et al., 1995; Odewahn, 1995) and support vector machines (SVM) galSVM (Huertas-Company et al., 2008; Huertas-Company et al., 2009, 2011).

The decision tree based technique of Random forests has been deployed quite frequently in astronomy, for example, in Breiman (2001). They were also found to be the most effective at identifying transient features in Pan-STARRS imaging (Wright et al., 2015). Here they are using pixel data and not survey catalogue data (comma delimited text files containing features of galaxies). Random forests were also used for the identification and classification of Galactic filamentary structures (Riccio et al., 2016), again using pixel data. Miller et al. (2015) used Random forests for the inference of stellar parameters using a training set of 9000 spectra for which stellar parameters were already known.

Wndcharm, a tool originally developed for use in biological image analysis, has been employed several times in astronomy (Orlov et al., 2008). It calculates over a thousand different features from images and then classifies test images into pre-defined classes. It uses the Fisher discriminant (Bishop, 2006), a supervised algorithm, to identify the features that result in the most accurate classification performance. Kuminski and Shamir (2016) used Wndcharm to classify $\sim$3,000,000 SDSS galaxies as spiral or elliptical.

Combining Galaxy Zoo classifications with supervised machine learning algorithms has become a more prominent research area (Banerji et al., 2010). In this case, a neural network, a multi-layer perceptron (MLP), classified galaxies into one of three classes: early types, spirals and point sources/artifacts. The training set consisted of features such as parameters based on colours and profile fitting, and a second dataset using features based on adaptive moments, and a combination of the two. More recently Wright et al. (2017) combines classifications from citizen scientists of the Zooniverse Supernova Hunters project with those from a Convolutional Neural Network (CNN) to identify supernovae (a transient, short-lived object) in data from Pan-STARRS1.

Perhaps the most exciting development in the machine learning field, certainly in the domain of image processing and understanding, is the development of the convolutional neural network (CNN) LeCun et al. (1989), and in particular its combination with larger training sets and graphical processing units (GPUs). Krizhevsky et al. (2012) won the Computer Science ImageNet competition in 2012. Krizhevsky's results were a significant improvement on the previous year's results. In subsequent years nearly all entrants were a variant of a convolutional network using GPUs and the machine learning community refocused their attention back to multi-layer neural networks under the overarching name of Deep Learning (Goodfellow et al., 2016). Significant further improvements to the state-of-the-art to the convolutional net architecture (CNNs) such as object detection (Redmon et al., 2016; Ren et al., 2015), in image segmentation (Long et al., 2015). Classification performance on specific tasks is now rivalling or outperforming humans (He et al., 2016). CNNs were used in the recent famous AlphaGo experiments to construct a representation of the position on the game board (Silver et al., 2016).

The first appearance of CNNs in astronomy was the winning entry to the GalaxyZoo Kaggle competition in 2015 which classified images of galaxies from the SDSS survey (Dieleman et al., 2015a). Dieleman based his solution on the network architecture by Krizhevsky et al. (2012) combining it with Goodfellow et al. (2013). He also modified the training data by clipping the galaxy image and rotating them to enable rotation invariance. The CNN associated the same galaxy at different rotations with the same target label i.e. making rotation irrelevant. The Galaxy Zoo classifications for the training set consisted of 32 weighted fractions (related to the decision tree). The CNN predicted these 32 values for each galaxy and compared them to those from Galaxy Zoo. Since this time better methods and CNN architectures have been developed (He et al., 2016), it is not clear whether translating these new models, as Dieleman did, will achieve better results.

Of direct relevance to my thesis is the use of a CNN to classify CANDELS galaxies into five morphological types (Huertas-Company et al., 2015). The approach they used was the technique developed by (Dieleman et al., 2015a). The training set used with the ConvNet relied on 8000 classifications made by 65 professional astronomers of galaxies in the GOODS-South field (Kartaltepe et al., 2015). The CNN was able to classify the remainder of CANDELS into five morphological types: disk, spheroid, peculiar/irregular, point source/compact and unclassifiable. These five classification types represent a fairly limited set compared to previous work by Dieleman et al. (2015a). One of the classifications was 'unclassifiable', so only four types were identifying previously recognised morphology. However, the unclassifiable class is interesting as the CNN effectively learnt to distinguish galaxies in the CANDELS dataset that human classifiers considered to be unclassifiable. Even though these unclassifiable objects had a wide variation of morphology the CNN was very effective at highlighting these objects. This is something that CAS has great difficulty with as the morphologies can be so varied.

The CNN is becoming much more prominent in astronomy and given its successes in other domains we can expect to see its use in many more applications in the future. Another recent application is to classify radio galaxies in archival data from the Very Large Array(VLA), Faint mages of the Radio Sky at Twenty-Centimeters (FIRST) survey (Aniyan and Thorat, 2017).

There are other configurations of deep neural networks developed in the field of computer science. For example, recurrent neural networks (RNN) including Long Short-Term Memory networks (LSTM) variants (Gers et al., 1999). These are used to model sequences of data and have most recently found use in image & video captioning (Johnson et al., 2016; You et al., 2016) – typically combining convolutional networks to model images and RNNs to model the captions. Also, text and language modelling (Sutskever et al., 2014) and speech recognition (Graves et al., 2013). These breakthrough techniques have been under-represented in application to astronomy, but one recent paper used an RNN to classify supernovae (Charnock and Moss, 2017). The sequence used as input were observational time and filter fluxes. The network successfully learnt light curves and could classify different types of supernovae with an accuracy of over 90%. They found that the results were sensitive to training set size. However, they were using 10,000 examples, which is considered to be a small training set for these kinds of networks. In addition, one-dimensional CNNs were used to classify light curves to detect exoplanets (Shallue and Vanderburg, 2018).

These techniques all employ supervised learning. Supervised learning has the disadvantage that it requires labelled input data, and so is limited in its potential for completely automated data analysis and exploration of large data sets. The upfront human acquired classifications drive the process and a supervised algorithm, such as the CNN, cannot classify objects outside of these pre-defined labels used in the training process. If an additional classification is required then the galaxies must be manually re-classified using e.g. Galaxy Zoo, and then the model must be rebuilt.

If classifying spectra or time-domain data such as light curves then LSTM networks or 1d CNNs have proven to be very effective. It also clear that the current state-of-the-art for automated classification of images is the CNN. However, to work effectively there needs to be enough labelled data. If training a CNN from scratch then hundreds of thousands of images can be necessary. However, if there is limited labelled data then one path that has not been used yet in astronomy is transfer learning. This technique has been very successful in other domains (Goodfellow et al., 2016). It involves using a pre-trained CNN and 'fine-tuning' the higher layers with available labelled data. For example, if I wished to classify a particular type of galaxy in CANDELS for which I have very few labelled galaxies, I could use a pre-trained CNN such as the one by Dieleman et al. (2015b) trained on SDSS images and then re-train the higher layers on the smaller number of CANDELS galaxies. Transfer learning works because the previously trained neural network has learnt features that are relevant to other optical image datasets. It would be worthwhile for the astronomical community if a pre-trained CNN was made available to the broader community.

The main advantage of these deep learning models is their complexity and their ability to learn features from the data that are most effective at predicting the labels in the training sets. The major disadvantage is that they require enough data in the initial training in order to reach an effective classification rate. However, transfer learning may be used to mitigate this problem if an alternative large training set is available in addition to the data of interest.

Perhaps the major disadvantage of CNNs is that they cannot yet quantify galaxy components in the manner that structural decomposition techniques are able to. It is possible that in the future CNNs could be adapted for this purpose. Another application that they are not suited to is general data mining and investigation of a dataset for the unknown. All galaxies will be classified based on the labelled data available. If there are populations of interesting galaxies that do not fit the labelled training data then the CNN will classify the objects depending on how similar they are to existing classifications. The extent to which they would be classified as 'unclassifiable' if such a group exists in the training data depends on the similarity of the galaxies to that class.

1.2.3.2 Unsupervised techniques

An alternative approach is the use of unsupervised machine learning algorithms. These algorithms enable exploratory data analysis and eliminate the need for human intervention (e.g. pre-labelling).

Unsupervised techniques enable us to identify the underlying structure within data. The underlying structure can be modelled and used to identify similar structure in new data. However, if there are no prior classifications how does one evaluate the results? What is the underlying meaning of the identified structure? To answer these questions a person needs to inspect the output of unsupervised machine learning algorithms. The advantage is that the output provided by these algorithms, such as a list of groups of data, or the identification of data that is similar, is a much smaller and considerably easier task than analysing the source data directly.

The potential for using unsupervised learning in application to astronomy has been recognised for over two decades. Most notably in the analysis of catalogue data to perform star and galaxy separation (Miller and Coe, 1996; Naim et al., 1997). It has also seen particular use in the estimation of photometric redshifts (Geach, 2012; Way and Klose, 2012; Carrasco Kind and Brunner, 2014), and object classification from photometry or spectroscopy data (D'Abrusco et al., 2012; in der Au et al., 2012; Fustes et al., 2013).

Another application has been finding galaxy clusters (over-densities of galaxies) (Ascaso et al., 2012). Ascaso's paper develops a technique they call 'Bayesian Cluster Finder' (BCF) to find galaxy clusters using catalogue data only. The type of data the tool uses consists of features from catalogue data including photo-z, galaxy x and y positions and magnitudes. No pixel data was used.

Unsupervised learning has also been used to search for anomalies, in particular, outliers in SDSS galaxy spectra (Baron and Poznanski, 2016). This work used an unsupervised configuration of the Random Forest algorithm to analyse two million galaxy spectra from the SDSS. They examined the 400 galaxies with the highest outlier score. The list included galaxy lenses, close galaxy pairs, galaxies with supernovae and others with unusual gas kinematics. They claim that the majority of the outlier galaxies had not been previously reported.

The previous applications of unsupervised algorithms have been to tabulated catalogue data and to spectra (detected flux values for the wavelengths of a galaxy's light). Much less work has been done to apply unsupervised algorithms to images directly.

Examples of techniques that claim to be unsupervised analysis of images include work by Schutter and Shamir (2015) presenting computer vision techniques to identify galaxy types (see also Banerji et al., 2010). This approach required an existing catalogue of galaxy images that are sorted by class at the input stage, which is pre-labelling and therefore a supervised process.

Other work by Shamir (2012) developed an outlier detection technique to detect peculiar galaxies among a training set consisting of a single clean morphological type. The technique trains unsupervised algorithms on a pre-labelled and collated training set. Shamir et al. (2013) used a pre-defined training set with supervised and unsupervised algorithms to classify galaxy mergers. Shamir and Wallin (2014) combined supervised and unsupervised techniques in an outlier technique to identify peculiar galaxy pairs in 400,000 SDSS images.

These examples all incorporate, in some part, unsupervised algorithms to classify images of galaxies. However, they all use the collation of a training dataset by pre-labelling galaxies. A completely unsupervised machine learning technique that can be applied to survey images without this upfront effort is arguably yet to be proven.

CNNs and their variants have been amazingly successful when used in a supervised setting to classify images, segment images and identify objects. When labelled data sets are available these techniques have delivered many state-of-the-art results. However, applying deep learning algorithms, neural nets in an unsupervised way is still an open area of research. More recently Generative Adversarial Networks (Goodfellow et al., 2014, GANs) a generative model mostly used to generate false, but realistic images have been employed by Schawinski et al. (2017) to denoise images of galaxies with much higher performance than simple deconvolution. However, this technique is not capable of grouping galaxies into classes. Another a variant of a GAN called Information Maximizing GAN (Chen et al., 2016, InfoGAN) claims to be able to group images of digits, but is untested on broader datasets. Other types of algorithm focus on something called self-supervised learning (Noroozi and Favaro, 2016). Self-supervised learning is a hybrid technique that uses supervised learning algorithm on a target variable created automatically from the data itself. For example, Noroozi and Favaro (2016) use a training set consisting of images split into tiles, and pairs of tiles are switched at random. The supervised algorithm then learns to correctly re-position the tiles. The correct positioning of the tiles is the supervised training signal. The network learns the latent structure of images by solving these ‘jigsaw puzzles’. The use of this technique in astronomy could be limited as the size of the tiles are quite large. These unsupervised algorithms are not in widespread use outside of machine learning research labs and do not reach the performance of equivalent supervised algorithms such as classification using a CNN.

One area that uses where neural networks could be used is for the purpose of learning a representation of images (Bengio et al., 2013). Particular types of neural network called Autoencoders and Convolutional Autoencoders can be used to reduce the dimensionality of data by learning a non-linear mapping of an image to itself. This technique will not identify different clusters or groups of images, but the output can be used as features, or as a proxy, for an image which can be used as input to other algorithms. Coates et al. (2011) used autoencoders for unsupervised representation learning, for use with supervised algorithms, but found the representations they created were not as effective as other techniques.

1.3 Research Questions

Existing techniques using unsupervised machine learning to group or classify galaxies all use a pre-labelled training set of images. A completely unsupervised machine learning technique that can be applied to explore imaging surveys without an upfront classification effort has not be proven. Therefore, I investigate the following three research questions:

- Can a technique using unsupervised machine learning algorithms segment survey images and identify galaxies?
- Can such a technique identify similar galaxies and categorise them into groups?
- Can the technique create a catalogue of galaxies found in a survey which facilitates astronomical research?

1.4 Contributions

- I have established a novel technical framework using completely unsupervised machine learning methods that can identify and segment galaxies in survey images without an upfront classification effort.
- I have evaluated the technique and established that it can successfully identify similar galaxies and categorise them into groups.
- I have created a novel catalogue of CANDELS galaxies (CANDELS is defined in Chapter 2) created as a result of applying the technique to automatically analyse the five CANDELS fields.

- The introduction and demonstration, to the field of astronomy, of the power spectrum feature as a successful method of image representation when used in combination with unsupervised machine learning. The power spectrum feature is defined in Chapter 3.

1.5 Publications

During the course of this work three peer reviewed documents were published.

1. Hocking, A., et al., 2015. Unsupervised Image Analysis & Galaxy Categorisation in Multi-Wavelength Hubble Space Telescope Images in European Conference on Machine Learning (ECML 2015) Doctoral Consortium.
2. Hocking, A., et al., 2017. Mining Hubble Space Telescope Images in Proceedings of the International Joint Conference on Neural Networks (IJCNN 2017).
3. Hocking, A., et al., 2018. An automatic taxonomy of galaxy morphology using unsupervised machine learning in Monthly Notices of the Royal Astronomical Society (MNRAS) Volume 473, Issue 1, 1st January 2018.

1.6 Thesis outline

In Chapter 2, I describe the data sets used to perform the experiments in this work. I describe how and where the survey data was acquired, the telescopes, the limits of the surveys, and how the data are represented in the data files used for the experiments.

In Chapter 3, I describe the pixel representations and algorithms that I considered for use in the technique. I look at the strengths and weaknesses of each algorithm and apply them to test data to identify their different characteristics. In the latter part of the chapter I discuss the options and methods for segmenting images, locating objects and representing galaxies.

In Chapter 4, I describe the technique in full, which algorithms I chose to perform experiments. I provide a complete description of the technical framework for its use to categorise galaxies and segment images. The final section of the chapter shows the initial results of segmenting survey images.

In Chapter 5, I describe the experiments performed to identify which pixel representations, algorithms and hyper-parameters that have the best performance. Also, I show how well the different pixel representations can identify strong-lensing artefacts in Hubble Frontier Field data.

In Chapter 6, I demonstrate the technique by applying the technical framework to *Hubble Space Telescope (HST)* image data. I then apply the technique to the five *HST* CANDELS fields and produce a catalogue for approximately 60,000 sources.

In Chapter 7, I summarise the conclusions arising from the experimental results, ponder over new questions, and provide suggestions for improving the technique. Finally, I describe the potential use of the technique to analyse Large Synoptic Survey Telescope images to localise electromagnetic counterparts of gravitational wave detections.

Appendix A contains descriptions of image analysis techniques that provide more details for the pixel representations described in Chapter 3.

Appendix B contains the description of the adjusted rand index clustering evaluation technique used in Chapter 5.

Chapter 2

Data

2.1 Introduction

The primary results presented in this thesis use data from the two surveys produced by the *Hubble Space Telescope*: the Frontier Fields (FF) (Bullock, 2012; J. Lotz, 2014) and the Cosmic Assembly Near-Infrared Legacy Survey (CANDELS) (Koekemoer et al., 2011). In this chapter I describe the *Hubble Space Telescope (HST)* and the two surveys.

To evaluate the output of the technique I use two high-quality and independent peer-reviewed sources of data the 3D-*HST* (Skelton et al., 2014) and Galaxy Zoo classifications (Simmons et al., 2016a). These data sources are used in Chapter 6 to identify the astronomical meaning and relevance of the groups of similar galaxies identified by the technique.

A goal of this thesis is to establish methods that use unsupervised machine learning that group similar galaxies based on pixel data alone. Therefore, the method uses no photometry or redshift data. If necessary, redshift and photometric catalogues may be used to analyse the results of the method by matching the catalogue created by the technique with external sources such as the *3D-HST* photometric catalogues of Skelton et al. (2014).

2.2 *Hubble Space Telescope (HST)*

2.2.1 Telescope design, Cameras and Filters

Human vision detects light at wavelengths from 390nm to 700nm. Astronomers typically refer to any telescope that operates at or around these wavelengths as an ‘optical telescope’. The *HST* is an optical telescope as its cameras and filters operate at 250-1700nm approximately (ultraviolet to the near-infrared).

Launched into low-Earth orbit in 1990, the observatory includes three cameras, two spectrographs, and a 2.4-metre telescope of a Ritchey–Chrétien (RC) design. Unlike older reflecting telescope designs, which use one or more parabolic mirrors, it has primary and secondary hyperbolic mirrors resulting in less spherical aberration and less coma at the edges of a wide-field-of-view (Wynne, 1968). Also, the RC design enables a long focal length resulting in high magnification, making it ideal for imaging small objects. The secondary mirror is held in place by four thin metal vanes which cause the cross-like diffraction patterns seen in observations of bright stars. Most professional telescopes employ an RC design including the mid-Infrared *Spitzer Space Telescope*, the Subaru and the dual 10 metre Keck Observatory telescopes on Mauna Kea, Hawaii.

The *HST* resides in a low orbit circling the Earth approximately every 95 minutes. It uses a high precision guidance sensor and up to three (of six) gyroscopes to maintain its position while observations are in progress. Despite the significant maintenance issues and expense involved in using an orbital telescope the benefits over a similar sized ground-based optical telescope are enormous. For example, ground-based telescopes must observe light that passes through the atmosphere. The atmosphere absorbs large sections of the ultraviolet and infrared spectrum, only allowing the band from 350nm to 760nm to pass-through. Also, distortions occur due to ‘seeing’ which is the blurring of the point spread function (PSF) as a result of distortions in the direction of incoming wave-fronts caused by lots of different refractive cells in the atmosphere. The fluctuations in humidity and weather contribute significantly to the noise and variability of ‘seeing’. These factors limit a telescope’s resolving power.

The *HST* has three cameras: the Advanced Camera for Surveys (ACS), the Wide Field Camera (WFC3) and the NICMOS camera which was superceded by the WFC3. Each camera has filters that can be used to image broad and narrow bands of light in the electromagnetic spectrum. The

ACS is designed for deep, wide-field imaging from the near ultraviolet to the optical. The WFC3 is the newest camera on the satellite. It has two channels, UVIS 200nm to 1000nm, and an IR channel operating between 900nm to 1700nm.

2.2.2 Data Processing

The Space Telescope Science Institute (STScI) has an extensive data pipeline to process the data retrieved from the telescope. Each camera requires broadly the same steps to convert the downlinked data into images and the process consists of removing issues to do with the CCD cameras such as:

- Pixels vary in size caused by manufacturing flaws and the natural limits of manufacturing precision.
- The geometry of the cameras on the telescope has a tilted focal plane causing geometric distortions in the images causing , for example, an elongation of the field-of-view.
- The filters contain small random variations.

These distortions to the pixel grid must be corrected in order to extract the highest quality images. The STScI has performed extensive work to accurately map these issues and provide 2D lookup tables that cover each of the CCD detectors in the cameras that can be applied to correct for the distortions. A software package, called DrizzlePac [1], is provided to perform this activity. Also, an additional software package, called CALACS, is available to perform the standard optical telescope calibrations and image pre-processing such as applying corrections for:

- Flat-fielding to remove anomalies as result of the variations in pixels' sensitivity to light. All pixels on the CCD have a natural variation in how sensitive they are to detections. This variation must be measured and corrected.
- Bias correction - measures and corrects for the variation caused by noise and interference in the electronics as the data is read off the sensor. It involves reading an image of the CCD with 0 seconds of exposure, ensuring any noise is the result of reading the data off the sensor.

[1] http://www.stsci.edu/hst/HST_overview/drizzlepac

- Dark correction - identifies the output produced by each pixel when there is no signal. This process accounts for issues such as hot pixels and small natural variations in the pixel detections when there is no signal.

There are extensive data handbooks for each camera which describe the calibration and corrective processes[2,3].

In terms of the effect on morphology and morphological classification performance many of these issues are not relevant. For example, the natural variation of pixels in the camera are very small and therefore the impact across an entire galaxy of size 40 pixels or more is negligible. Perhaps the most difficult problem is the geometric distortions caused by the tilted focal plane which can, to a limited degree, elongate the images of the galaxies. Although, the amount of geometric distortion is unlikely to be detected by current morphological classification techniques.

Survey images are typically mosaics of observations taken by the telescope. The field-of-view of the telescope is not large enough to image large areas of the sky. Combining observations into mosaics allow individual images to contain larger areas. This process requires accurate image alignment using the DrizzlePac software. The surveys described in the next section are mosaics of many observations.

The image data is stored in a file format called Flexible Image Transport System (FITS files). The FITS format contains metadata that allows astronomers to map a pixel position in the image to a co-ordinate in the sky and vice-versa. Figure 2.1 shows an example of the image data provided by the STScI after the CALACS and DrizzlePac software was applied. The Figure contains a section of an image from two FITS files. This image is part of the Frontier Fields survey. The individual images are monochromatic grey scale consisting of a 2D array of data, with each pixel represented by a 32bit float value. By imaging the galaxies through different filters we can seek to discern and measure the physical processes occurring within the galaxies.

2.3 Colour and Morphology

It is important to recognise that colour and morphology are considered as two separate properties. Astronomers use both to help them understand galaxies. Colour is usually defined as the

[2] http://www.stsci.edu/hst/HST_overview/documents/datahandbook

[3] http://www.stsci.edu/hst/acs/documents/tirs/tir1402.pdf

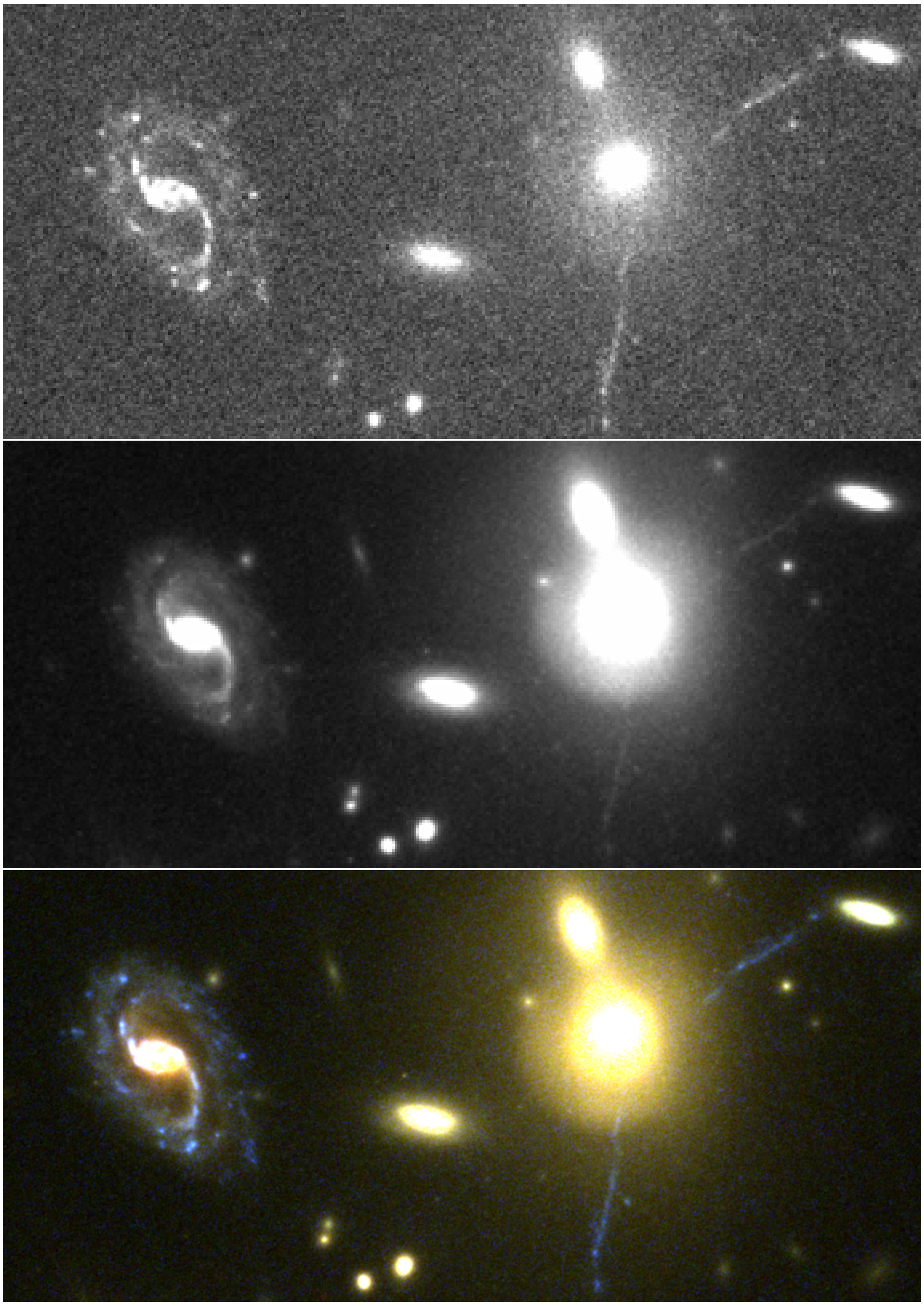


FIGURE 2.1: The first image at the top is a subsection of a FITS image file. The image was taken using the ACS camera and the F435W filter over many orbits of Hubble. The F435W is a broad band filter centred around 435nm, which is near the shortest wavelength available to Hubble. The second image is also produced using many orbits of observations with the ACS camera, but this time using the F814W filter. The bottom image is a composite created by using the STIFF tool (Bertin, 2012) to combine the image data from F435, F606W and F814W representing the R, G and B channels.

difference in magnitude between two wavebands. Observing galaxies at different wavelengths, or in different colours, may lead to differences in observed morphology. This is because the colour reflects differences in the underlying physical processes. For example, the most massive stars emit light predominantly in blue and ultraviolet wavelengths and have relatively short lifespans, compared to less massive stars that emit predominantly in red and last for billions of years. Therefore if we observe a galaxy that is emitting a significant amount of blue light, we know that it is a star-forming galaxy, as it still contains short-lived blue stars. On the other hand, if we observe a galaxy that is predominantly red in colour, we know that there are no short-lived blue stars and therefore there is little star formation. Spiral galaxies generally have blue spiral arms, indicating that they are star-forming, whereas elliptical galaxies are usually yellow, or red, suggesting that they are passive (Roberts, 1963; Kennicutt Jr, 1998). Therefore a spiral galaxy observed only at longer wavelengths is likely to be classified differently from a spiral observed at short wavelengths. A galaxy observed at blue wavelengths may also appear more clumpy than one observed at longer wavelengths, as the blue light will highlight localised areas of star formation that are not apparent at longer wavelengths.

Traditionally galaxies were classified based on their morphology in a single waveband and did not make use of colour, but modern morphological classifications may make use of composite images made from several bands. Colour may also be used independently to classify galaxies as star-forming or non star-forming (Baldry et al., 2004). This approach has lead to the identification of two main types of galaxies, star-forming (the blue cloud) and passive (the red sequence), plus a smaller group of galaxies with particularly high star formation rates (sometimes referred to as starbursting galaxies). Some galaxies are also found to lie at intermediate colours, indicating lower star formation rates, in what is sometimes referred to as the green valley (Schawinski et al., 2014). Colour can also be used to identify galaxies with unusual properties (e.g. green peas, Cardamone et al. (2009)).

2.4 Frontier Fields

The first data set used in my thesis is the *(HST)* Frontier Fields (FF) images (Bullock, 2012). This dataset consists of a deep field observing program designed to image six strong lensing galaxy clusters and six parallel blank fields (that contain a normal distribution of galaxies e.g. there are no strong lensing clusters). The galaxies behind the strong lensing galaxy clusters experience magnification, with small regions magnified by factors of up to 100. The parallel and cluster observations have a depth of 29 ABmag[4] at 5σ[5] with small regions magnified up to 33 ABmag (Lotz et al., 2017). The parallel blank fields enable statistical analysis of the distribution of distant galaxies brighter than 29 ABmag. A large enough blank area enables biases associated with the variance of the distribution of galaxies along every sightline to be measured (Lotz et al., 2017).

Gravitational lensing is a consequence of Einstein's Theory of General Relativity. It is observed where the path of photons is deflected as a result of the curvature of spacetime caused by the presence of matter (Einstein, 1936; Zwicky, 1937). Strong lensing is the observation of magnified background galaxies and, possibly, multiple images of the same galaxies, as result of the presence of a sufficiently high mass density (Ellis, 2010).

The six strong lensing clusters are Abell2744, MACSJ0416.1-2403, MACS1149, MACS0717, Abell S1063, and Abell 370. Each of the massive galaxy clusters, and the six parallel blank fields located near to the clusters, were imaged using both the WFC3 (F105W, F125W, F140W, F160W) and the ACS (F435W, F606W, F814W). 70 orbits were used to complete the imaging of each field. The Frontier Fields images were released starting in April 2014 and completed in 2016. The galaxy clusters were selected for the survey because of their predicted lensing strength and observability with other observatories such as the Atacama Large Millimeter/submillimeter Array (ALMA) FF Survey (González-López et al., 2017, ALMA-FF), telescopes on Mauna Kea, Hawaii and the Spitzer Space Telescope.

Figure 2.2 shows the Frontier Fields galaxy cluster called Abell 2744. This image was created by combining the data from images taken using the ACS camera and filters F435W, F606W and F814W. The colour image was produced using the STIFF tool (Bertin, 2012).

[4] AB magnitude (ABmag) is the logarithm of the spectral flux density. Spectral flux density is the rate at which energy is transferred by electromagnetic radiation through a surface (or a virtual surface), per unit surface area per unit wavelength.

[5] To be accepted as a detection the signal must be at least 5 times the standard deviation (σ) of the background level. So the faintest objects that can be detected in this survey at 5σ have a magnitude of 29 ABmag.

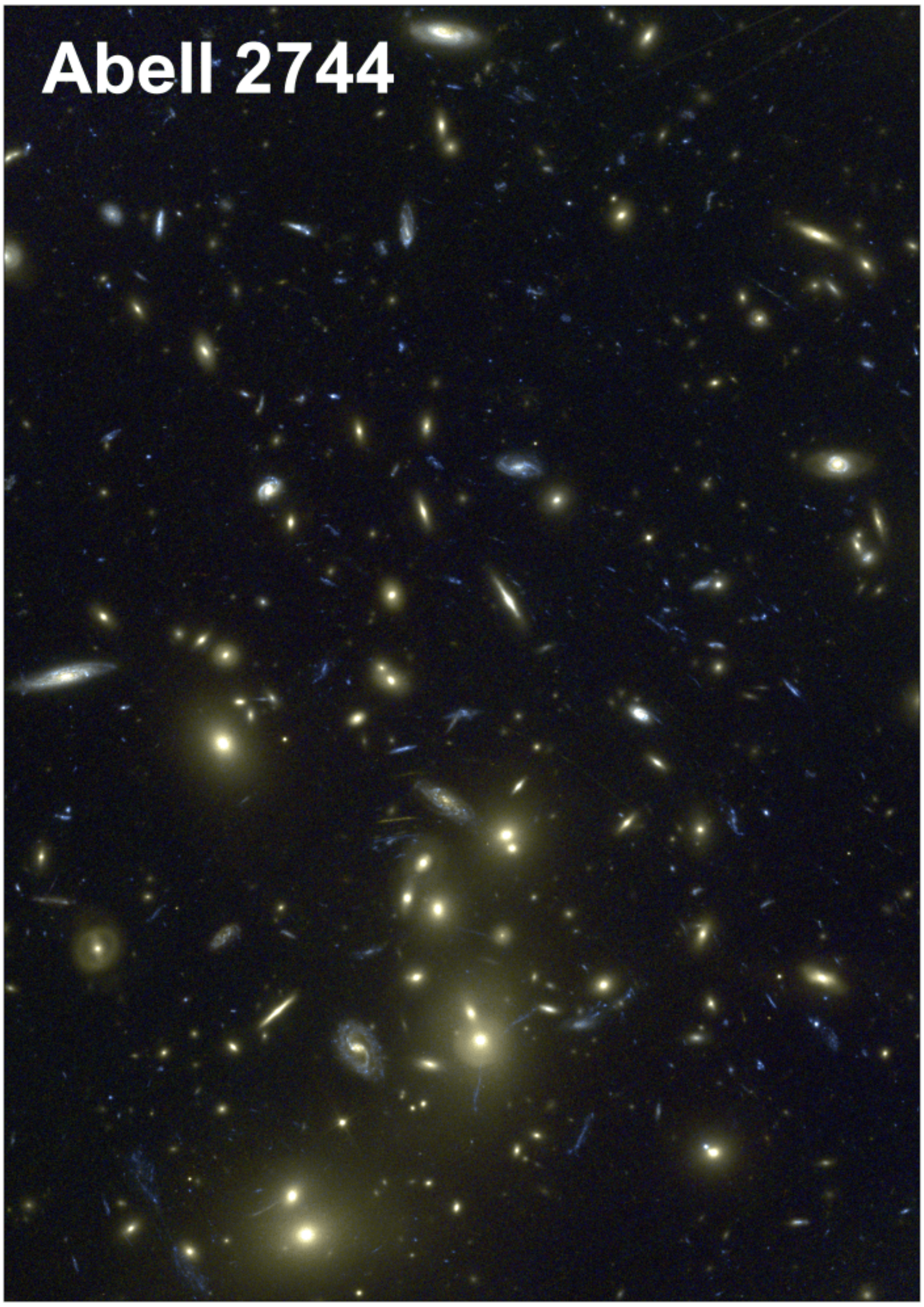


FIGURE 2.2: This is an RGB composite image of the *HST* Frontier Field Abell 2744 (90″× 130″). The red, green and blue channels correspond to the F814W, F606W and F435W bands.

2.5 Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS)

The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (Grogin et al., 2011; Koekemoer et al., 2011, CANDELS) is the largest *HST* survey designed to document the evolution of galaxies out to $z \approx 8$. The survey consists of the WFC3 optical and infrared (WFC3/UVIS/IR) and the ACS optical imaging of five extragalactic survey fields. There are two tiers: a 'deep' survey to at least four orbit effective depth of 27.8 ABmag in F160W over $\sim$125 arcmin^2 in the two Great Observatories Origins Deep Survey (GOODS) fields of GOODS-South and GOODS-North. The second tier is a wider and shallower survey to two orbit effective depth of 27 ABmag in F160W covering $\sim$800 arcmin^2 of COSMOS, EGS and UKIDSS/UDS and flanking areas of GOODS-South. For all five fields I used version 1.0 release of the data[6], selecting the filters F160W and F814W as they provide the most complete coverage across all five fields. The F814W images were projected onto the same grid as the F160W (0.06$''$ per pixel). Figure 2.3 shows the GOODS-North field and the mosaic created by combining the individual *HST* observations can be seen.

2.6 Data and Catalogues Used for Evaluation

The technique described in this thesis identifies morphologically and photometrically similar galaxies and sorts them into groups on that basis. In order to evaluate the outcome I use two independent peer reviewed sources of data that present detailed data and classifications of the galaxies in CANDELS:

1. Skelton's 3D-*HST* CANDELS photometry and photo-z catalogues (Skelton et al., 2014).

   These catalogues are based on an analysis of images from 147 data sets covering all five CANDELS fields, at wavelengths 0.3-8 um. They provide AB magnitudes, together with 1σ uncertainties, and photometric redshifts determined using EAZY. The photometric redshifts were compared to available spectroscopic redshifts reaching normalized median absolute deviation scatter of <2.7% with fewer than 5% significant outliers.

2. Galaxy Zoo classifications for three of the five CANDELS fields (Simmons et al., 2016b).

[6] https://archive.stsci.edu/prepds/candels/

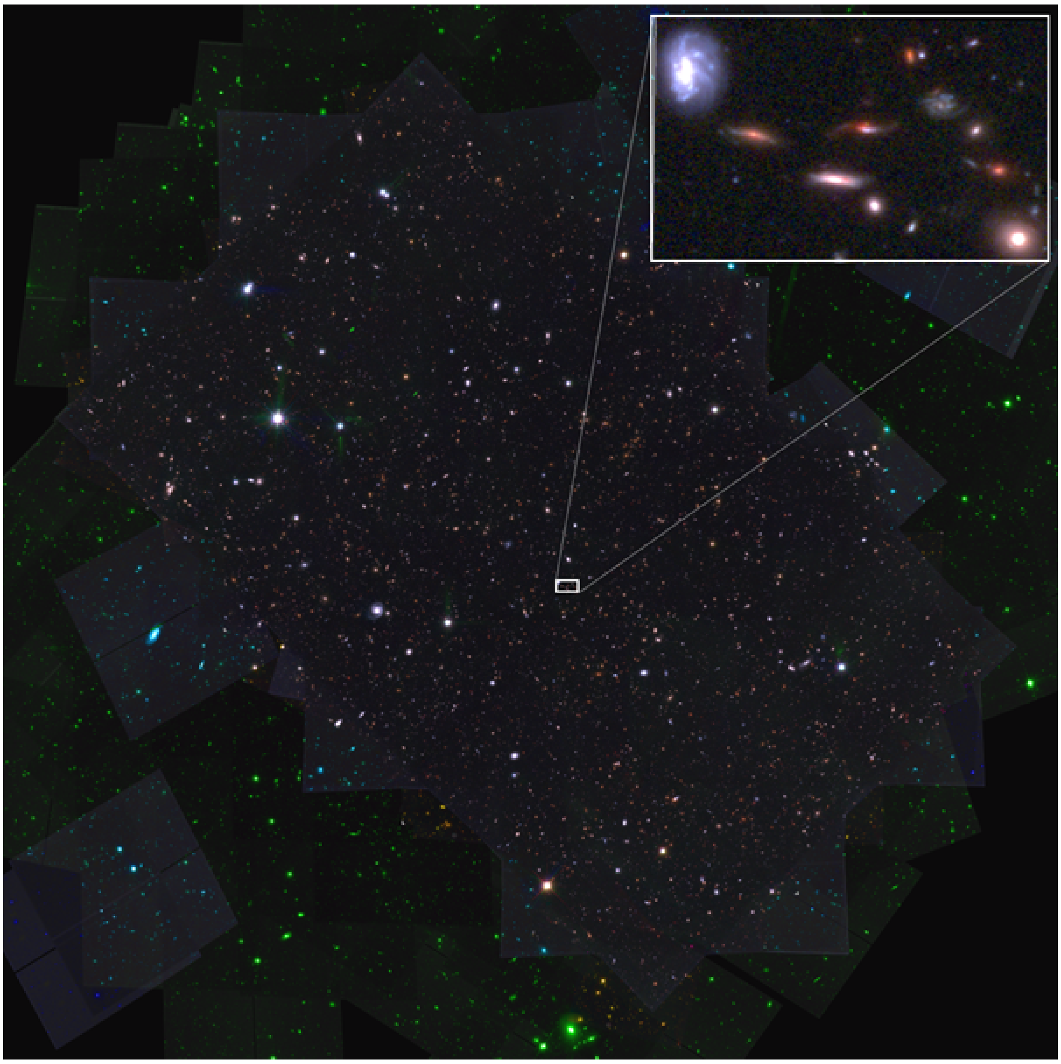


FIGURE 2.3: This is a composite RGB image of the CANDELS data for GOODS-North. It is a combination of data from observations using filters F160W, F814W and F606W. The STScI team combine the *HST* observations into a mosaic. The field is 158 arcmin^2. The individual square observations ($\sim$ 2.2arcmin^2) can be seen, and where the image is green or blue showing the lack of coverage in all filters.

Citizen scientists classified 48,000 galaxies by answering morphology related questions from a decision tree (see Figure 2.4). Each galaxy received an average of 40 classifications. The answers to the questions are converted into weighted voted fractions. Each classifier is given an initial weighting based on their performance separating stars from extended sources. The weighting scheme then down-weights the classifications from classifiers who are error-prone or who regularly diverge from the consensus.

The catalogue provides one of five 'clean' classifications for those galaxies where there is high confidence.

Depth corrections are provided for the GOODS-South deep field. The depth corrections were identified by having the citizen scientist classify a subsample of GOODS-South galaxy images of similar depth to the COSMOS and UDS fields. This is in addition to the 'deep' images already available of the same galaxies. The depth corrections are identified by comparing the classifications of the 'shallow' and 'deep' images.

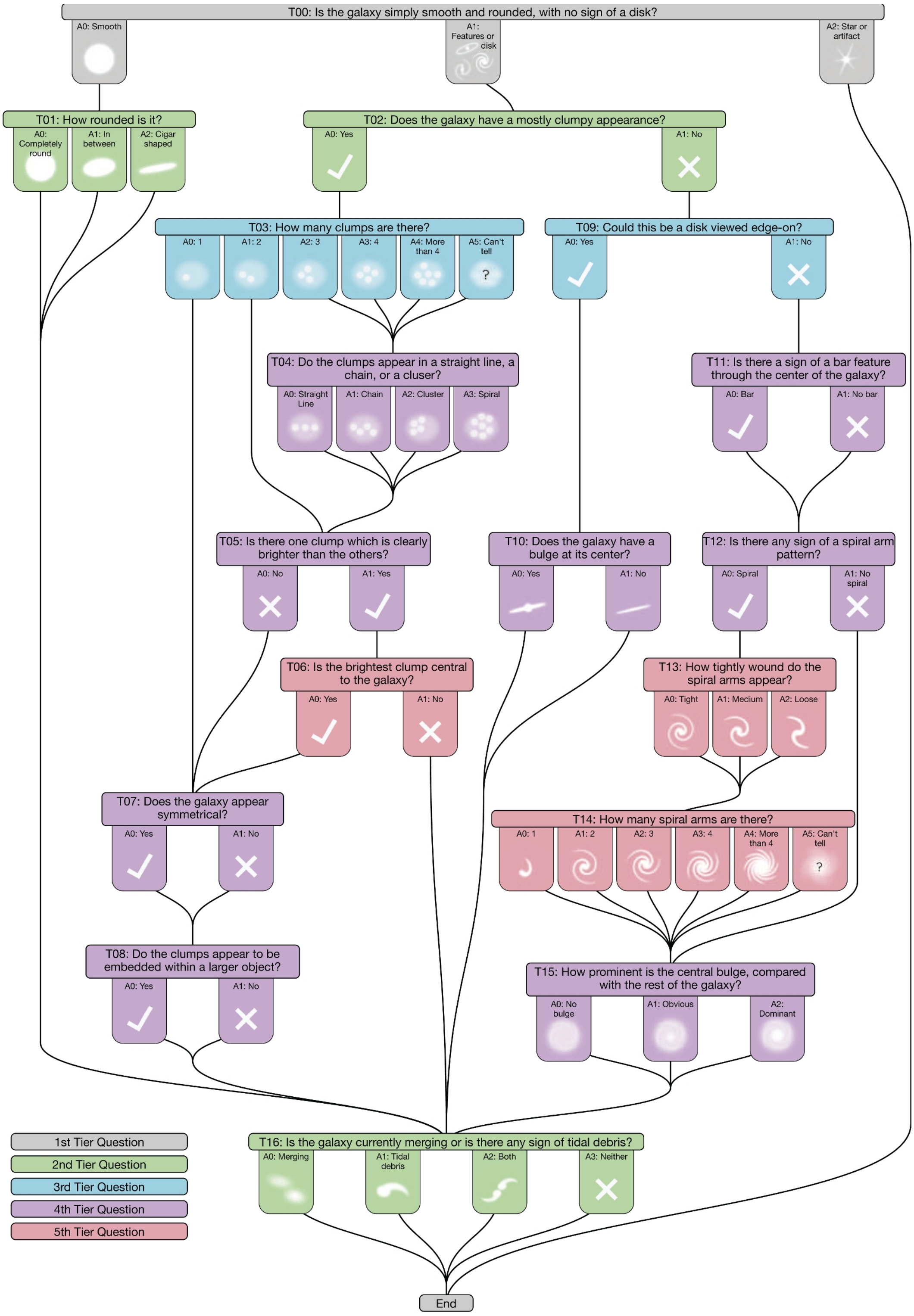

T00: Is the galaxy simply smooth and rounded, with no sign of a disk?
A0: Smooth
A1: Features or disk
A2: Star or artifact
T01: How rounded is it?
A0: Completely round
A1: In between
A2: Cigar shaped
T02: Does the galaxy have a mostly clumpy appearance?
A0: Yes
A1: No
T03: How many clumps are there?
A0: 1
A1: 2
A2: 3
A3: 4
A4: More than 4
A5: Can't tell
?
T09: Could this be a disk viewed edge-on?
A0: Yes
A1: No
T04: Do the clumps appear in a straight line, a chain, or a cluser?
A0: Straight Line
A1: Chain
A2: Cluster
A3: Spiral
T11: Is there a sign of a bar feature through the center of the galaxy?
A0: Bar
A1: No bar
T05: Is there one clump which is clearly brighter than the others?
A0: No
A1: Yes
T10: Does the galaxy have a bulge at its center?
A0: Yes
A1: No
T12: Is there any sign of a spiral arm pattern?
A0: Spiral
A1: No spiral
T06: Is the brightest clump central to the galaxy?
A0: Yes
A1: No
T13: How tightly wound do the spiral arms appear?
A0: Tight
A1: Medium
A2: Loose
T07: Does the galaxy appear symmetrical?
A0: Yes
A1: No
T14: How many spiral arms are there?
A0: 1
A1: 2
A2: 3
A3: 4
A4: More than 4
A5: Can't tell
?
T08: Do the clumps appear to be embedded within a larger object?
A0: Yes
A1: No
T15: How prominent is the central bulge, compared with the rest of the galaxy?
A0: No bulge
A1: Obvious
A2: Dominant
1st Tier Question
2nd Tier Question
3rd Tier Question
4th Tier Question
5th Tier Question
T16: Is the galaxy currently merging or is there any sign of tidal debris?
A0: Merging
A1: Tidal debris
A2: Both
A3: Neither
End

FIGURE 2.4: This is the decision tree used to guide citizen scientists through the classification process for the CANDELS dataset. Each citizen scientist starts at question T00 (top). Each galaxy is classified multiple times and each classification is provided by a different person. The answers for each question are consolidated into weighted consensus classification. Credit: Galaxy Zoo & Simmons et al. (2016a)

Chapter 3

Machine Learning Methods Considered

Sections 3.3.2, 3.3.3 and 3.4.2.1 of this chapter were published in Hocking et al., 2018 *Monthly Notices of the Royal Astronomical Society*, 473, 1108

3.1 Introduction

So far I have described the objective of using an unsupervised machine learning technique to segment images, detect objects and then group (sort) the objects based on their visual similarity. I now consider and evaluate options in terms of how to represent image data and which unsupervised algorithms have the most useful characteristics to fulfil the outlined goals. I must also keep in mind that we need to use the technique to categorise new images in very large data sets.

Unsupervised algorithms have been used in astronomy to identify objects. For example, Shamir and Wallin (2014) use a technique to identify peculiar galaxy pairs by preparing a pre-collated training set of positive and negative examples of galaxy mergers. They use a technique developed in Shamir (2012); Shamir et al. (2013) based on the Wndchrm[1] tool (Orlov et al., 2008). The positive examples were images of true galaxy mergers and the negative examples were images of galaxies that were clearly not mergers. Although no labels were used by their technique, the acquisition and use of the training set is clear supervision. I, however, seek a purely unsupervised technique whereby no pre-collation of a training set is required. Instead the technique

[1] https://github.com/wndcharm/wndcharm

must analyse an existing area of sky, identify latent structure in that data and group similar objects. The model will then be used to identify similar objects in new images. The key questions I consider to develop this technique are:

1. What is the best method to represent pixel data: Should pixel intensities be used as a feature or should they be converted a pre-designed feature? Is the rotation of objects a concern and should I solve this by using a rotationally invariant feature? Also, would dimensionality reduction be useful by reducing the number of features? It is not generally possible to predict which features perform the best and therefore it is common practice in machine learning to test a series of features and see which gives the best results.
2. Clustering Algorithms: Is it important to use an algorithm that can automatically identify the correct number of groups? Which algorithms have the highest performance? And which algorithms produce the best result in segmentation and object sorting?
3. Object detection method: Galaxies often appear very close together or overlapping, therefore how should this issue of contamination be resolved? Also, galaxies typically appear in many orientations, therefore, how can the technique be made agnostic to rotation and orientation?

I start by considering how to best represent pixel data in order to segment images.

3.2 Representation of Pixel Data

Galaxies of the same type can appear in any orientation. Therefore a primary requirement is that the model be robust to galaxy rotation, orientation and scale. Many feature representations when applied to images of galaxies would result in objects with the same orientation and size being grouped together, for example, a group of thin, vertically aligned edge-on galaxies being grouped together, and horizontally aligned edge-on galaxies in a separate group. This is not desirable as morphologically they are the same type of galaxy and therefore we want them to be grouped together regardless of their angle of rotation. The final choice of representation must be free of this bias.

3.2.1 Features

I now select and evaluate feature representations and their invariant properties for their use in an unsupervised technique. The following options were selected for consideration because of their simplicity, in the case of pixel data, or their effectiveness when applied to normal imaging. Apart from the power spectrum, all the following features considered are widely used in image processing.

1. Pixel intensities.

   This is the simplest option. By extracting sub-images I could concatenate images from different filters. The problem with using pixel intensities directly is that they are not robust to rotation. It is also a high-dimensional as each pixel is a single dimension feature which could be potentially difficult to work with for some algorithms (Domingos, 2012) and slow to process and dataset sizes are potentially very large. In supervised machine learning pixel data is almost never used directly unless used with CNNs (see Chapter 1) where the algorithm learns a representation of the data.

2. Histograms of Oriented Gradients (HoG) (Dalal and Triggs, 2005, HoG).

   HoG has been successfully used for tasks such as detecting people in images, character recognition in images, and face recognition (Dalal et al., 2006; Newell and Griffin, 2011; Déniz et al., 2011). The HoG feature works by calculating the gradient changes in an image and representing them in a histogram of directions and magnitudes. Sharp gradient changes in images usually coincide with edges and corners of objects. Appendix A.1 explains how image gradients are calculated and Figure 3.1 shows the HoG feature for an image of an elliptical galaxy. In terms of galaxies this would capture the gradients of central bulges and spiral arms etc. However, this feature is not rotation invariant. Histograms representing the gradients will be very different for similar galaxy features at different angles. HoG would be useful for symmetrical central bulges and elliptical galaxies but is not not effective for asymmetrical objects such as spiral arms, edge-on galaxies and irregular galaxies.

3. Rotationally Invariant Feature Transform (RIFT) in Lazebnik et al. (2005)

   In many ways this feature is simply a rotationally invariant version of the HoG feature. It calculates the orientation of the gradient with respect to concentric rings radiating from the centre. Rotation invariance occurs as the orientation is measured at each point relative to the direction pointing outward from the centre instead of the horizontal and vertical directions in HoG. This is also distinct from other feature descriptors such as Scale Invariant Feature Transform (Lowe, 2004, SIFT) and AKAZE (Alcantarilla et al., 2012, AKAZE) which require the identification of a single dominant orientation for the image, calculated by summing the orientations in the horizontal and vertical directions. There is no dominant orientation for a galaxy, unlike cars for instance, therefore, the SIFT and AKAZE approach is unlikely to lead to effective rotation invariance for galaxies. Parameter options for RIFT include the number of rings and the number of histogram bins to produce a vector for each image. The size of the vector is the product of the number of rings and the number of bins. Figure 3.2 shows how this feature vector is calculated.

4. Spin intensity images in Lazebnik et al. (2005) This is a two dimensional histogram that encodes the distribution of pixel intensity values in a rotationally robust way. Once again it uses concentric rings radiating from the centre. A histogram is created of the pixel values on each ring. The histograms are concatenated to form the vector representing the image patch. The key difference between this feature and RIFT is that it only uses the intensity pixel values and does not use the gradient changes in the image. The output histogram consists of bins of the intensity of the pixels only. Figure 3.3 shows how the spin intensity feature is calculated.

5. Pixel intensity power spectrum

   This is a feature not seen used with unsupervised machine learning algorithms. I calculate the power spectrum of the pixel intensities in an image patch by: calculating the 2D discrete Fourier transform (DFT) of an image patch, using a Fast Fourier Transform algorithm (Cooley and Tukey, 1965, FFT). I then multiply by the conjugate of the output of the DFT, rearrange the zero frequency component to the centre of the spectrum (to the centre of the 2D matrix) and then calculate its azimuthally averaged radial profile. Figure 3.4 shows how this is done for a picture of an elliptical galaxy. Appendix A.3 provides a description of how the 2D discrete fourier transform, (DFT) and 2D power spectrum is calculated and provides an description of the aspects of imaging, such as spatial frequency, that are identified by using this method.

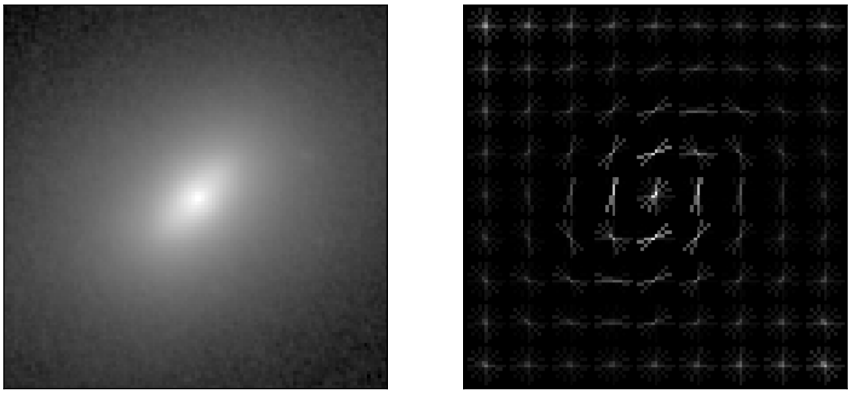


FIGURE 3.1: A visualisation of the Histogram of Oriented Gradients (HoG) feature descriptor. On the left is an image of a galaxy and on the right is the visualisation of the HoG feature. The image is segregated into 12×12 pixel cells and within each of these cells a histogram with 8 radial bins is created. The image to the right shows the histograms in each cell with the orientations and magnitudes of each bin represented by short lines. The size of the vector output is the number of radial bins by the number of pixel cells in the image.

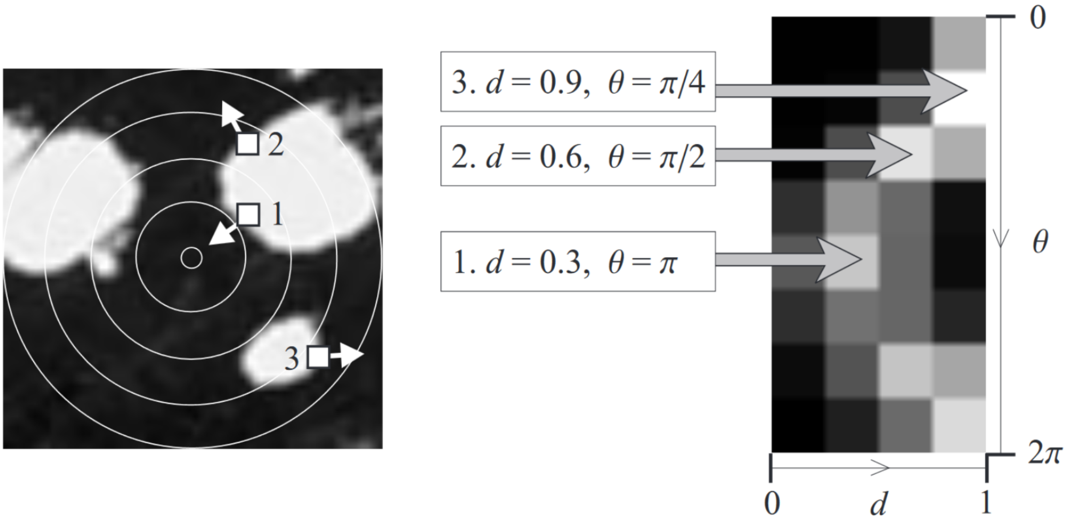

2
1
3
3. $d = 0.9,\ \theta = \pi/4$
2. $d = 0.6,\ \theta = \pi/2$
1. $d = 0.3,\ \theta = \pi$
0
θ
2π
0
d
1

FIGURE 3.2: The RIFT feature descriptor. A patch is divided into concentric rings of equal width and a gradient orientation histogram is computed within each ring. The orientation is measured relative to the direction pointing away from the centre. This maintains rotation invariance. A typical configuration is four rings and eight histogram orientations resulting in a 32 dimensional feature descriptor. Three sample points in the normalised patch (left) map to three different locations in the descriptor (right). d is the distance from the centre, and θ is the direction of the gradient. The magnitude of the gradient is indicated by brightness of the histogram cell. Credit: Lazebnik et al. (2005)

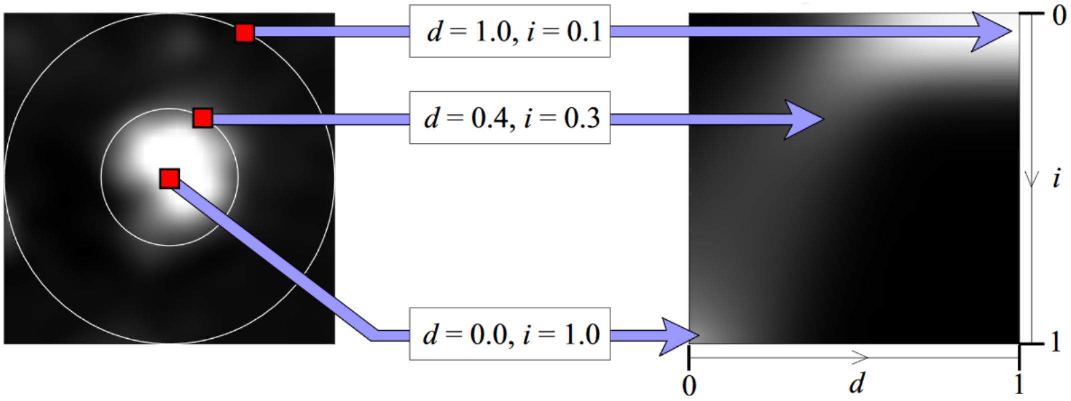


FIGURE 3.3: The spin intensity feature descriptor. This descriptor encodes the distribution of image brightness in the region of a centre point. d is the distance from the centre point and i is the intensity value forming a two dimension histogram (right). Each row of the histogram represents a histogram of the intensity value at a distance d from the centre of the patch. Three example sample points from the image patch (left) map to three locations in the descriptor (right). A typical configuration is to use ten bins for distance and ten bins for intensity values, resulting in a one-hundred dimension vector. Credit: Lazebnik et al. (2005)

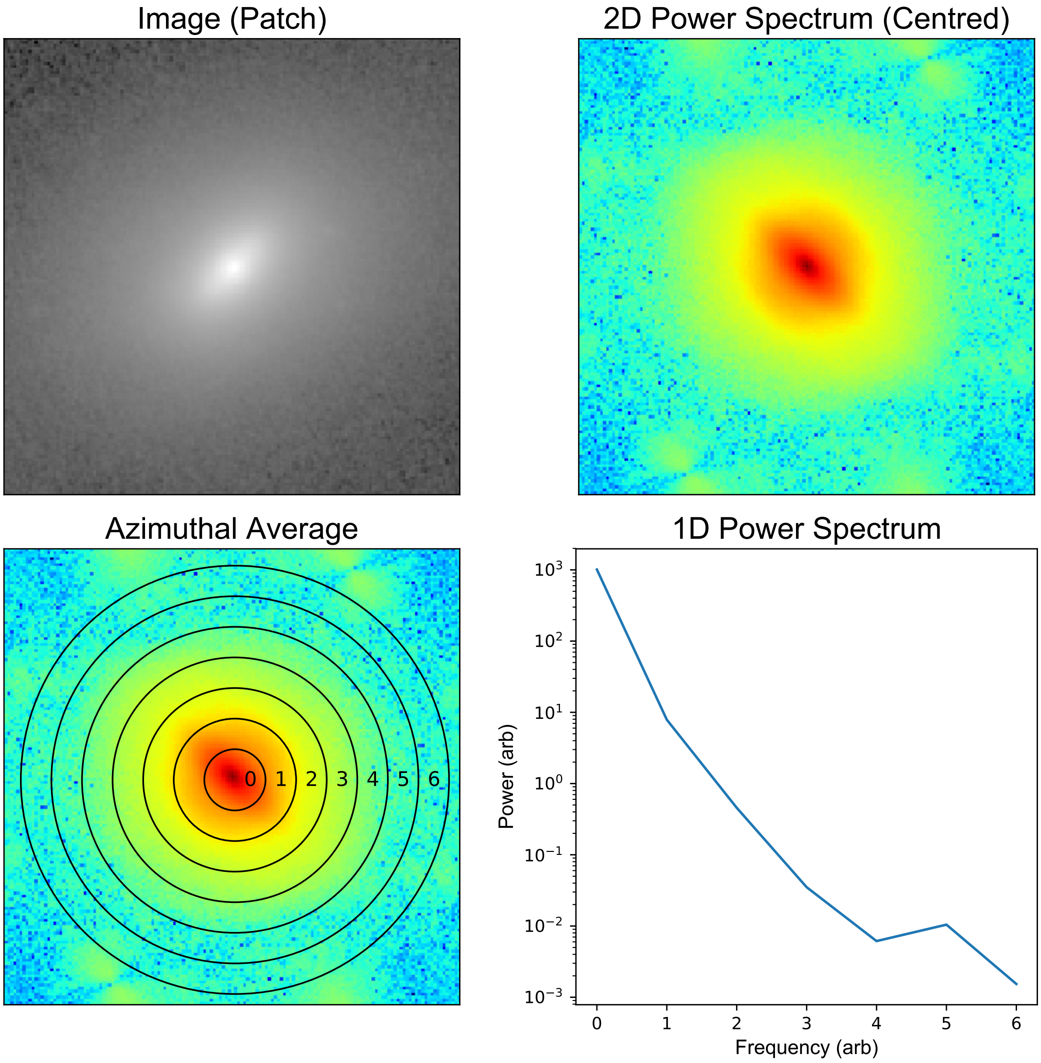


FIGURE 3.4: Power spectrum feature descriptor. The original image (patch) is shown (top left): this is an image of a galaxy, the 2D discrete fourier transform is calculated and the result is shown (top right), the average of the power values of the pixels within seven annular bins (bottom left), finally the averages form a 1D representation of the power (bottom right). This results in seven features (one for each annular bin) for this patch.

3.2.2 Key Point Detectors

A key point is also known as an interest point that are positions in an image that are highly distinctive, such as corners, areas of peak brightness and large gradient changes. These points can be found if an image is resized or transformed. The majority of key point detection methods use a concept called the scale-space of an image to help identify corners and key points and areas where scale-space shows significant changes (See Appendix A.2 for an explanation of scale space). This has been a popular technique in image processing to create features for machine learning algorithms. Many techniques exist that exploit scale space such as SIFT (Lowe, 2004), SURF (Bay et al., 2008), ORB (Rublee et al., 2011) and BRISK (Leutenegger et al., 2011) among others. The results of applying a key point detector, AKAZE (Alcantarilla, 2011; Alcantarilla et al., 2012) to a FITS image can be seen in Figure 3.5. This shows that AKAZE was somewhat effective at finding the peaks. The green points are the key points detected by AKAZE. So it is evident that the gradient changes in the image at the centre of galaxies were identified as key points. However, dim objects were left undetected. This is not surprising as the gradient changes within the image are much higher in the bright objects. It may be possible to adjust the scale space to capture more subtle changes within large images of individual galaxies however, it was not clear that this would be improve the results. In addition these key point detectors are not useful for segmenting images which is one of my requirements.

I choose not to use key point detectors because they only identify the brightest points of the brightest galaxies. They do not detect other useful points such as individual components of galaxies. Therefore, using key point detectors is very unlikely to produce useful results when grouping galaxies based on morphology and is of no use when segmenting galaxies.

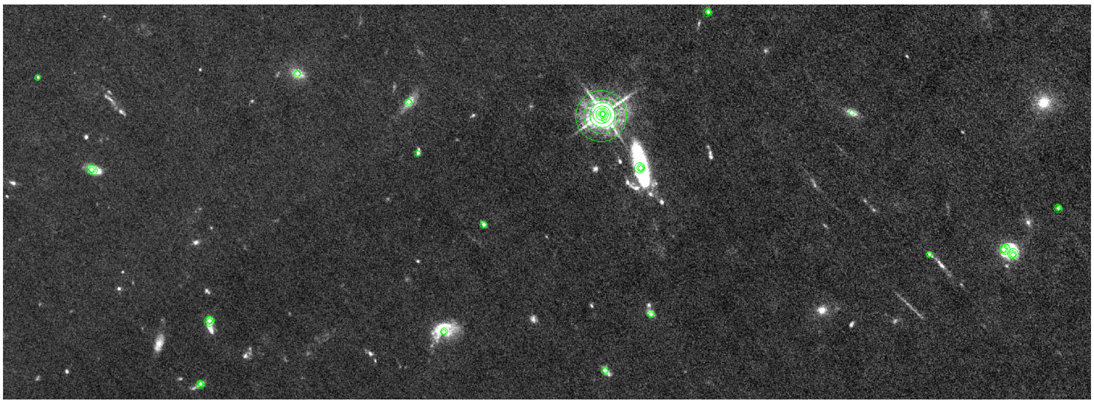


FIGURE 3.5: AKAZE Key point detection on a *HST* FITS grayscale image. The green points are the key points detected by AKAZE. The gradient changes in the image at the centre of galaxies were identified as key points. However, not all the galaxies were detected and there is usually only one key point per galaxy. It is possible that this method could be developed for the identification of sources but it shows little promise for classifying or clustering galaxies into groups. Therefore scale-space key point detectors are unlikely to be useful when applied to survey images. It maybe that they are more useful for individual images of large galaxies where enough detail is resolved for the key point detectors to identify galaxy components.

3.3 Unsupervised Algorithms

I have considered the following established clustering algorithms:

1. K-Means (Sculley, 2010)
2. Growing Neural Gas (GNG), toplogical, graph based algorithm (Fiser et al., 2012; Fritzke, 1995).
3. Hierachical Clustering (HC) (Hastie et al., 2009)
4. Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999)
5. DBScan (Ester et al., 1996; Birant and Kut, 2007)

These algorithms were chosen because they are all unsupervised but approach the problem of clustering using a variety of different commonly used approaches: NMF uses matrix decomposition, GNG is a graph based algorithm, DBScan uses density modelling and (HC) creates a hierarchical representation of data and K-Means is a baseline algorithm. One issue with clustering algorithms is that many require the number of clusters to be specified beforehand. An additional reason to consider the GNG and DBScan algorithms is their claimed ability to automatically identify the number of clusters in data.

I not that there are a large number of algorithms that were not included in my evaluation. These include probabilistic methods such as unsupervised Gaussian mixture modelling (Hastie et al., 2009), and finite mixture models (Figueiredo and Jain, 2002). Initial testing of finite mixture models showed it to be ineffective at finding clusters for high dimensional data.

3.3.1 K-Means

K-Means (Hastie et al., 2009) has been a very popular method of partitioning a data manifold for clustering purposes. K-Means clustering aims to partition m data points into K clusters so as to minimise the distances within each cluster. It starts performing an initial seeding step to randomly identify K data points to allocate as cluster centres. It then iteratively performs two steps: the first identifies which data points are closest to each k cluster centre, the second calculates the means for all the data points allocated to each cluster centre, this then becomes the cluster centre. The iterations continue until convergence and stabilisation of updates to the k cluster centres. This results in the data manifold/parameter space partitioned into Voronoi cells.

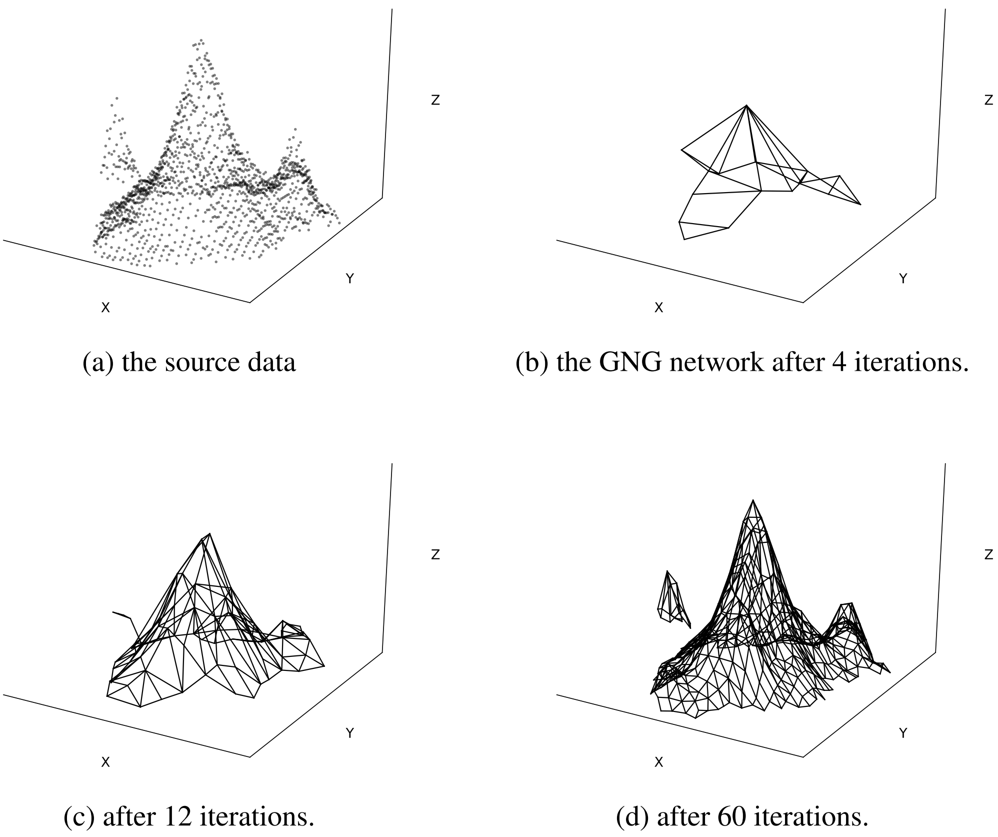
K-Means is a non-convex optimisation problem. Therefore the initial seeding of the clusters affects the clustering result. Typically the K-Means algorithm is run several times using random initialisation. The best result, identified by the mean intra-cluster distance, is then used as the clustering result.

There are several different formulations including K-Means++ to improve initial choice of the k clusters (Bachem et al., 2016). There are also versions intended to dramatically improve performance, for example, the use of mini-batch optimization whereby cluster centres are updated using a learning rate and samples drawn from mini-batches of the original data (Sculley, 2010).

3.3.2 Growing Neural Gas

The Growing Neural Gas (GNG) algorithm (Fritzke et al., 1995) creates a graph that represents the latent structure within data. The algorithm is used for the purposes of clustering and analysis of many types of data. GNG is applied to an $m \times n$ data matrix representing the input data that contains m rows, consisting of n-dimensional vectors called sample vectors. GNG identifies structure by iteratively growing a graph to map the data in the sample vector space. The graph consists of nodes connected by lines called edges. Each node has a position in the data space called a code vector. This is illustrated in Figure 3.6. The code vectors have the same dimensionality as the sample vectors in the data matrix. The algorithm starts by creating a graph of two nodes. Each node is initialised using a random sample from the data matrix. The graph grows and shrinks as the input data is processed (i.e. more samples are introduced). During this process the positions of the nodes evolve: the code vectors are updated to map the topology of the data and the graph splits to form disconnected sub graphs, each of which represents a cluster in the data space. The process continues until a stopping criterion has been met, such as a saturation value for the number of nodes within the graphs, or the processing time. In order to create a graph that accurately maps the input data it is common to process the input data multiple times. The learning steps of the algorithm are:

1. *Initialization* Create a graph with two nodes. Initialise the position of each node with the vector of values from a random sample vector $\mathbf{p}$ from the data matrix. Subsequently, samples are drawn at random from the data matrix and the following set of rules applied:



(a) the source data

(b) the GNG network after 4 iterations.

(c) after 12 iterations.

(d) after 60 iterations.

FIGURE 3.6: These four images show how the Growing Neural Gas (GNG) algorithm works to map and approximate data. (a) shows the sample data. The images (b), (c) and (d) show the progress of the GNG algorithm as it discovers and learns the structure of the data.

2. *Identify the two nodes nearest to the sample vector* For each node in the graph, the distance d between the sample vector $\mathbf{p}$ and the node's code vector $\mathbf{q}$ is calculated using the squared Euclidean distance. The two nodes (s_0, s_1) most similar to the sample vector (i.e. the two smallest values of d) are identified.

3. *Create and update edges* If an edge connecting s_0 and s_1 does not exist, create it. Set the 'age' of the edge connecting s_0 and s_1 to zero. Increment the age of all other edges connected to the nearest node s_0.

4. *Increase the 'error' of the nearest node* s_0 The 'error' is simply the squared Euclidean distance between a sample vector and nodes in the GNG: if the error is high then the GNG has not yet properly mapped the data space containing the sample vector. In this step the squared Euclidean distance between the input sample vector and s_0 is added to the local error of s_0.

5. *Move the nearest node s_0* Update the code vector of s_0 using equation 3.1. This step moves the nearest node s_0 'towards' the input sample vector $\mathbf{p}$. The ε_b parameter controls the size of the movement towards the input sample.

$$\Delta\mathbf{q}_{s_0} = \varepsilon_b(\mathbf{p} - \mathbf{q}_{s_0}) \tag{3.1}$$

6. *Move connecting nodes' neighbours* Using the same process as in the previous step but using the ε_n parameter to control the magnitude of the adjustment for nodes directly connected by an edge to s_0.

7. *Remove old edges and nodes* Remove all edges with an age greater than the maximum age A parameter. All nodes without edges are removed.

8. *Add a new node to the GNG graph* A new node is added to the graph after a fixed number (λ) of sample vectors have been processed. The new node is added at the midpoint between the node with the highest error and its connecting node. If multiple nodes are connected then the new node is positioned at the midpoint of the connecting nodes with the highest error. When a new node is added, the error of each node is reduced by α.

9. *Reduce all node error values* Reduce the error of each node in the GNG graph by a factor of β.

Fritzke et al. (1995) describes the parameters mentioned above in detail. The majority of the compute time is in step (ii); various attempts have been made to reduce the time taken (Fiser et al., 2012; Mendes et al., 2013). For a data matrix with few dimensions using various tree based methods to store GNG nodes works well and provides a significant performance increase over a brute force method to identify the nearest neighbours in the first step of the algorithm. However, as the dimensionality of the data increases the performance of the graph based methods decreases to become similar to that of the brute force method. I implemented a version of GNG that parallelises the brute force method of finding nearest neighbours as this provides the most flexibility.

3.3.3 Hierarchical Clustering

Hierarchical clustering (HC) (Hastie et al., 2009) involves a recursive process to form a hierarchical representation of a data set as a tree of clusters. One of the key benefits of HC is that it can

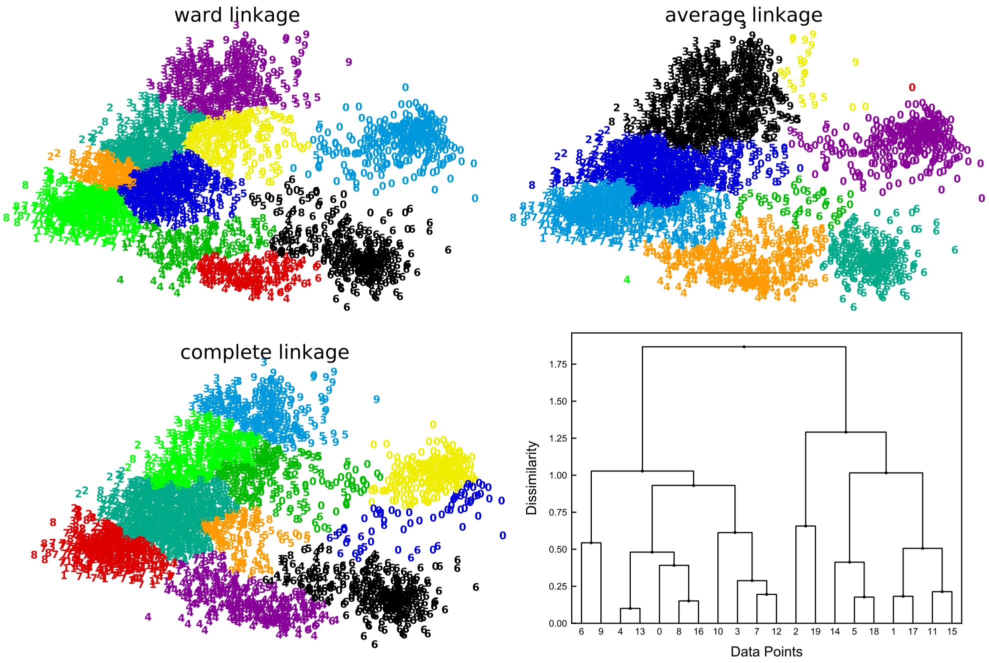


FIGURE 3.7: Top left, top right and bottom left show the identified clusters using agglomerative clustering on images of digits. Bottom right is a dendogram visualisation of the hierarchy identified by an agglomerative clustering process. The *x* axis identifes the data points. The *y* axis represents the degree of similarity. The lines identify the clusters the data points belong to, and the length of the lines are a measure of similarity. The root node of the dendogram is at the top. Credit: the code for the top left, top right and bottom left linkage plots from scikit-learn.org

produce uneven clusters, both in terms of their disparate sizes and separation in the parameter volume. In Figure 3.7 we can see the coloured clusters are not homogenous i.e. the shapes are uneven and the boundaries are non-linear. Whereas algorithms such as K-Means find homogenous clusters which is a bias of the algorithm rather than representing the inherent state of the data. Later in this Chapter I use test data with evenly shaped clusters, however, as you'll see in Section 3.4.1 the actual astronomy image data does not have an intrinsically even structure.

The identified clusters form a hierarchical representation of the input data, as illustrated in Figure 3.7. This agglomerative representation can be thought of as a tree structure where the leaves represent the individual input sample vectors from the data set. The process starts by merging pairs of leaves, using a measure of similarity to identify the most similar pair of leaves.

The pair with the closest proximity are merged into a new cluster (twig) that is added to the tree as a new parent node to the pair. The process continues by merging pairs of nodes at each level until a single node remains at the root of the tree. The final tree representation contains multiple 'levels' of nodes, with each node in a level representing a cluster. Each level can be considered a level of detail in a clustered representation of the data. The process of using a similarity measure to merge clusters is called linkage. I apply 'average' linkage which uses the chosen similarity measure to compare the centroids of the clusters at each level of the tree; a centroid is calculated by finding the average sample value within a cluster. After assessing the pairwise distance between all clusters in a level, clusters with the minimum linkage are merged, and the centroid of the merged cluster recalculated, ready for the next merging step as one moves up the hierarchy towards the single root.

There are a number of methods used to measure similarity between vectors, including Euclidean distance, Pearson correlation and cosine distance. After experimenting with these three types we found the best results were obtained using the Pearson correlation coefficient (see equation 3.2) and cosine similarity (see equation 3.3) measures,

$$r(\mathbf{p},\mathbf{q}) = \mathrm{cov}(\mathbf{p},\mathbf{q})\mathrm{var}(\mathbf{p})^{-0.5}\mathrm{var}(\mathbf{q})^{-0.5} \tag{3.2}$$

where r is the Pearson correlation between $\mathbf{p}$ and $\mathbf{q}$ (the code vectors from two GNG graph nodes) and

$$\cos(\theta) = \frac{\mathbf{p} \bullet \mathbf{q}}{||\mathbf{p}||\,||\mathbf{q}||} \tag{3.3}$$

the cosine similarity is the cosine of the angle between the two vectors.

Each node in the tree can be given a unique label and so the input data can be classified according to which node in the tree best describes it, at some desired 'level of detail' (the trivial example is that the 'root' by definition would label all of the data). In this work I am concerned with imaging data, and the algorithm described above can be used to label individual (or groups) of pixels in an image, therefore automatically segmenting and classifying them. Consider an image containing two different types of object: provided the data matrix captures the difference between these objects (be it morphology, colour, intensity, etc.), then the algorithm described above should automatically identify and label these two objects differently.

3.3.4 Non-Negative Matrix Factorization (NMF)

NMF was originally presented in application to image processing by Lee and Seung (1999). It is a matrix factorization technique that decomposes data into two low-rank non-negative factorising matrices. NMF has the constraint that the data and the factorising matrices must be non-negative. However, data can typically be translated by a constant factor if the data contains contains negative values. The technique identifies two matrix factors W and H whose product approximates X by iteratively minimising the root mean square residual between the input data X and $W \times H$. The algorithm requires two inputs, the data X, an n rows by m columns matrix, and k for the number of components. The size of the factors are: W is an n by k matrix and H is a k by m matrix. The factor matrices W and H are initialised to random values. The optimisation is non-convex and therefore, repeated runs maybe necessary to achieve optimum results and will result in different W and H factors.

This approximation problem is often formulated by choosing the Euclidean distance between X and WH for calculating the residual:

$$\min_{W,H \geq 0} \parallel X - WH \parallel^2 = \sum_{i,j} (X - WH)^2_{ij}$$

The resulting two matrices can be used for clustering purposes. Matrix factor W contains cluster centroids and H contains cluster membership indicators.

The non-negativity constraint of NMF implies that its use is only for applications where the data matrix is composed of non-negative elements. This constraint is realistic for many real world problems, and in particular for the domain we are applying these algorithms too: pixel intensities of images. Many data sets can be truncated or translated so all values are positive or zero.

Similar to other clustering algorithms, choosing k, the number of latent factors to identify, is not simple decision without knowledge of the underlying data. One can run the algorithm with different values to identify the optimum result.

3.3.5 DBScan

Another algorithm that claims to automatically identify the number of clusters is Density Based Spatial Clustering with Noise (DBSCAN) (Ester et al., 1996). This algorithm recognises clusters when the density of data points rises above a threshold level. Lower density areas are considered 'noise' and each data point within the cluster neighbourhood of a given radius has to contain at least a specified minimum number of data points (the density has to reach a threshold level). An advantage is that any distance function can be used to calculate proximity. The choice of distance function dictates the shape of the neighbourhood, for example, Manhattan distance in 2D space is rectangular. When visualising data, the Euclidean distance is typically used.

Given two parameters, n the number of minimum points in a neighbourhood for a point to be considered a core point and ε the maximum distance from the core point to be considered in the same neighbourhood. Clusters are discovered using the following process:

1. Find ε-neighbours of every data point. If there are more than n data points then a cluster is started. Otherwise the data points are considered noise.
2. Find connected components of core points. This establishes if core points are grouped into spatially separated clusters.
3. Assign non-core points to nearby clusters at a threshold distance of ε, or assign them as noise.

Useful benefits of this process are that the number of clusters is automatically defined by the number of dense data points above the threshold level. Also, the algorithm can identify uneven clusters shapes. Areas of low density are considered noise and therefore outliers are usually identifies as low density noise.

Although, the number of clusters does not need to be specified the user must still identify two parameter values: ε and n. Ester et al. (1996) offer heuristics to assist in their estimations. ε is dependent on the problem being solved (the distance) and n is the desired minimum cluster size.

A disadvantage of DBScan is that there can be only one combination of ε and n and therefore the algorithm assumes that densities of clusters are homogenous across a data sets. So it is not effective on data sets with large differences in densities as the two parameters can not be chosen for all clusters. In addition, the choice of neighbourhood distance function has considerable affect on the results and is tightly coupled to the choice of ε.

3.3.6 Visualisation Methods

In order to view the results of clustering algorithms I use two visualisation algorithms, a linear algorithm called Principle Components Analysis (Pearson, 1901, PCA) and a non-linear algorithm called t-distributed stochastic neighbour embedding (Maaten and Hinton, 2008, t-SNE) which is very effective for visualising high dimensional data.

PCA is as a linear dimensionality reduction technique. Data of high dimension can be visualised by projecting it into a two or three dimensional space. PCA aims to do this by removing correlated features and identifying the principle components of a data set. One method to calculate PCA is to first standardising the data by centring each feature to have a mean of zero. Calculate the covariance for the data set. The principle components are the eigenvectors of the covariance matrix ordered by their corresponding eigenvalues. The eigenvector with the highest eigenvalue is the first principle component. Typically a decision is made to retain the number of principle components that represent a threshold percentage of the information of the original dataset.

t-SNE is a non-linear method of dimensionality reduction. It converts high-dimensional Euclidean distances between data points into conditional probabilities that represent similarity. The idea is that this similarity should be the same when the data are projected into a lower dimensional space. In this paper Maaten and Hinton (2008) they use the Kullback-Leibler divergence[2] to measure the difference between the two distributions as a cost function. This is a non-convex optimisation procedure and therefore an optimisation method such as gradient descent can be used to minimise the cost function.

3.3.7 Motivation for algorithmic selection

I have reviewed a variety of unsupervised machine learning algorithms including an algorithm that performs topological mapping (GNG), another that uses matrix factorisation (NMF) and other standard clustering algorithms such as K-Means and hierarchical clustering. To motivate my choice of which algorithms to take forward I must consider performance, effectiveness and flexibility. Qualitatively we can see that each of these algorithms has benefits. For example, GNG and DBScan do not require a value to be set a-prori for the number of clusters to be found, and each algorithm claims to be able to automatically identify the number of clusters. Growing

[2] The Kullback-Leibler divergence (KL) is defined as: $KL(P||Q) = \sum_i P_i \log(\frac{P_i}{Q_i})$, where P is the true distribution and Q is a model distribution. The KL divergence measures the deviation of Q from P.

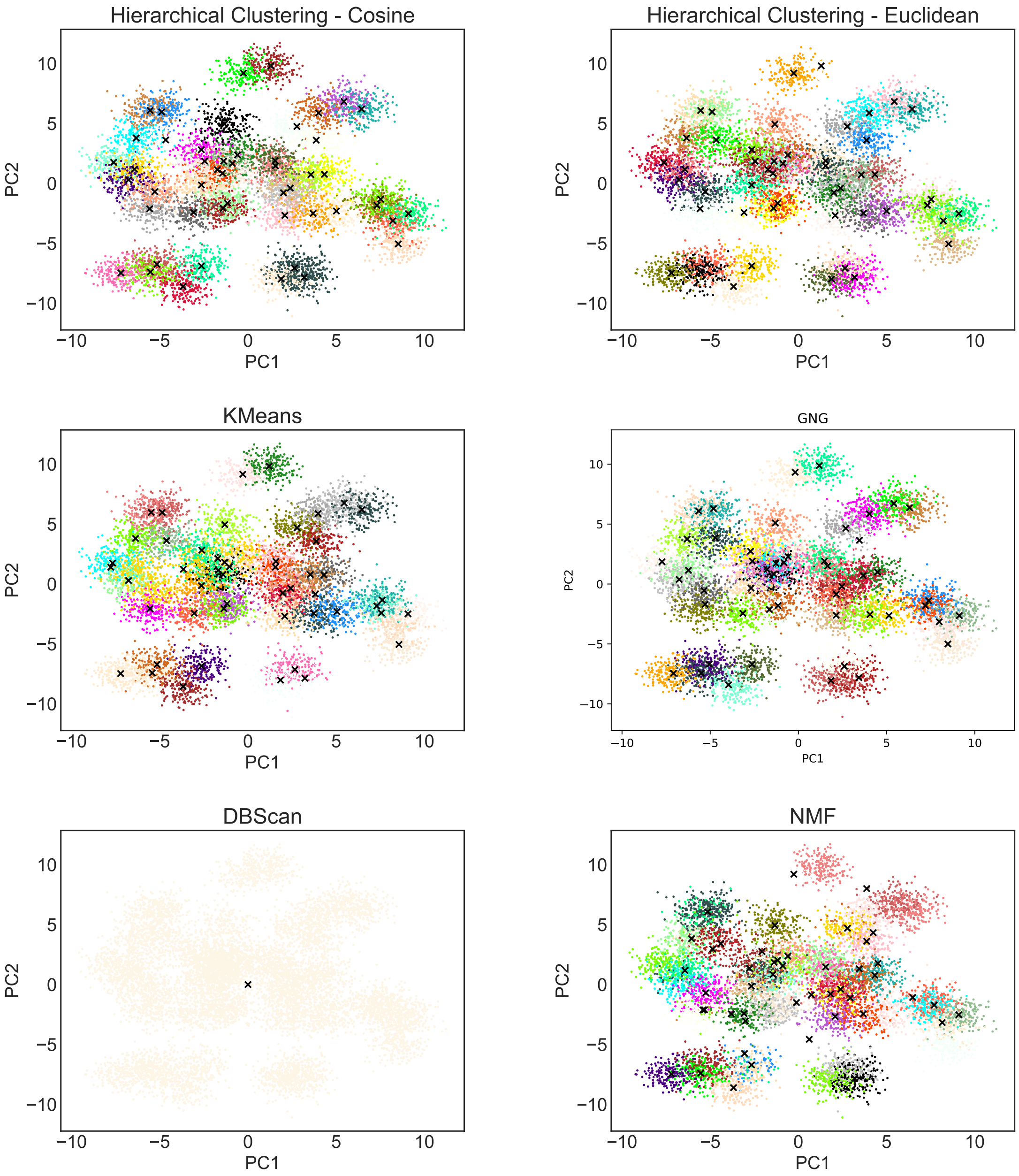


FIGURE 3.8: The algorithms were run on a set of high dimensional test data that consisted of 2000 features, with 50 spatially separated clusters. Each scatter plot shows the two principle components of applying PCA to the test data and the cluster centres identified by the algorithms were projected into the PCA space so they could be added to the plot. The retained variances for the first two principle components are 1.07%, 1.06%. The percentages are so low as the data has 2000 dimensions.

Neural Gas can topologically map data and identify uneven clusters. However, both DBScan and GNG have other parameters that must be set upfront. Hierarchical clustering allows for alternative dissimilarity measures to be used which is a useful property to be able to identify uneven clusters, however, its runtime performance is $O(n^3)$.

TABLE 3.1: These are the results of running the algorithms on four test data sets. The data sets differed only in number of features: 25, 250, 750, and 2000 features. This is to test how well the algorithms perform on data with high numbers of dimensions which is typical of image data. Each dataset contained 50 spatially separated, identically sized clusters created using random values from a Gaussian distribution with the same standard deviation. The performance measure used to identify how well the algorithms found the clusters is the normalised mutual information (NMI) score. A value of 1 shows the algorithm correctly identified the cluster membership of all data samples.

	NMI Scores			
Dimensionality / No. of Features	25	250	750	2000
Hierarchical Clustering (Cosine)	1.0	1.0	1.0	1.0
Hierarchical Clustering (Euclidean)	1.0	1.0	1.0	1.0
GNG	1.0	0.99	0.99	1.0
KMeans	1.0	1.0	0.99	0.87
NMF	0.80	0.98	0.98	0.78
DBScan	1.0	0.0	0.0	0.0

In order to inform the decision I test each algorithm on a toy dataset. This data consists of 10,000 samples, 50 Gaussian components. Figure 3.8 shows a visualisation of the results of processing the data with the six algorithms. I have projected the data into two dimensions using PCA. The algorithms were run on the original data set and the identified centroids were projected into the PCA space. The samples were given a colour according to their cluster assignments.

In order to evaluate the clustering ability of the algorithm I use the evaluation metric called Normalized Mutual Information (Vinh et al., 2010, NMI). NMI is the information theoretic measure called Mutual Information but adjusted to account for chance. Mutual Information values are higher for two clusterings with larger numbers of clusters irrespective of whether more information is shared. If I consider a toy example to demonstrate consisting of a data set with only four data points. I use NMI to calculate how similar two sets of cluster assignments are: (0,0,1,1) and (0,0,1,1), the NMI between these two cluster assignments is 1 i.e. they are considered to be the same. A second toy example is the cluster assignments (0,0,1,1) and (1,1,0,0), also resulting in an NMI of 1. They are also considered the same. This is correct as, although the cluster assignment identifiers are different, the data points are grouped together into the same clusters.

Table 3.1 shows the results of deploying the algorithms on the same toy data set, as shown in Figure 3.8, to establish basic accuracy clustering rates. We can see that Hierarchical Clustering, GNG and K-Means have the best results overall. DBScan showed excellent result for lower dimensionality.

I tested DBScan in accordance with the guidelines provided by Ester et al. (1996) specifying the *minimumpoints* parameter to be double the number of features. I ran the algorithm multiple times on the same data with increasing values for the ε parameter. I found that the ε parameter was effective in a narrow range and that the algorithm was very good at identifying clusters when the parameter was set optimally. However, any deviation away from a narrow range led to poor results. When the number of features was increased from 25 the algorithm did not find any clusters. The results were not robust to the parameter ε. The ε parameter is more difficult to set for data with a large number of features which is a typical property of image data.

The optimum methods in terms of flexibility and performance appear to be the Growing Neural Gas and hierarchical clustering algorithms. The reasons are that GNG can topologically map the data and reduce the number of data points by mapping rows to graph nodes. Also the parameters that it requires appear to be robust, the defaults appear useful for a variety of data. The ability of hierarchical clustering to use a multitude of dissimilarity measures to identify uneven clusters to cut the graph. So I am only using one part of the GNG algorithm, the toplogical mapping part and disregard the connected component graph cut.

I chose to explore the use of K-Means, Hierarchical Clustering and Growing Neural Gas. I use GNG to reduce the number of samples where each node or vertex of the graph represent one or more samples and I use hierarchical clustering to cut the graph of nodes.

I note however the test data used to evaluate the algorithms represents, to some extent, an idealised situation. This is because the clusters in the data are formed of well separated Gaussian components. These clusters should be very simple for the algorithms to identify. This approach was taken to evaluate whether the algorithms were worth trying on the astronomy data. In Section 3.4.1 I find that the astronomy image data is organised as large amorphous structures. An additional more relevant test would include data organised to more closely represent astronomy image data.

3.4 Integration of Methods

I have considered which features and algorithms could be the most effective and should be included in experiments. This answers the first two questions presented at the beginning of this Chapter. I now answer the third question of how to identify objects and how to represent them in order to group them by using unsupervised machine learning methods. The first step of this process is to segment images and I then consider how the segmented images can be used to represent objects.

3.4.1 Unsupervised image segmentation

Image segmentation is the process of identifying and partitioning images into regions. The result of a successful segmentation process is an image where regions perhaps representing objects are identified and their boundaries clearly marked. In the case of survey images in astronomy a segmented image would show the regions containing objects. Segmentations can be performed at different levels. For example, a segmentation process could identify the common components within objects i.e. the common areas across galaxies such as passive and active regions.

A supervised process would use a training set where the different areas to be segmented are marked so that the supervised algorithms can learn to map pixel data, or some representation of pixel data in the training set to a target value. However, using unsupervised machine learning we do not have a training set, or labelled target. Instead we rely on the technique to cluster areas that have similar underlying structure. Once the pixel data has been clustered a segmented image is produced by marking each pixel by a unique identifier representing the cluster that the algorithm identifies the pixel data belongs to. This can be completed at different levels of detail. For example, by using a windowing technique we can cluster small windows to produce a finer segmented image, or larger windows to produce a coarser segmentation.

In order to investigate and visualise astronomy data I used the power spectrum feature described in the previous section and extracted data from a *HST* Frontier Field image of galaxy cluster Abell 2744. Figure 3.9 shows the data visualised using a linear dimensionality reduction technique, PCA, and then using a non-linear dimensionality reduction called t-SNE. We can see that the data is inherently non linear. In addition, the data contains extreme outliers which are the central, brightest parts of galaxies which are in some cases orders of magnitude brighter than a typical area of the image. By applying the natural log we can see that the PCA plot has an improved appearance as a result of the data rescaling. This is important to note as, with the exception of DBScan, the algorithms we have previously discussed are not robust to outliers. The behaviour of clustering algorithms will be to partition the outliers as clusters, and the other data points as an individual cluster. The application of the natural log allows the clustering algorithms to cut the main body of the data points into clusters.

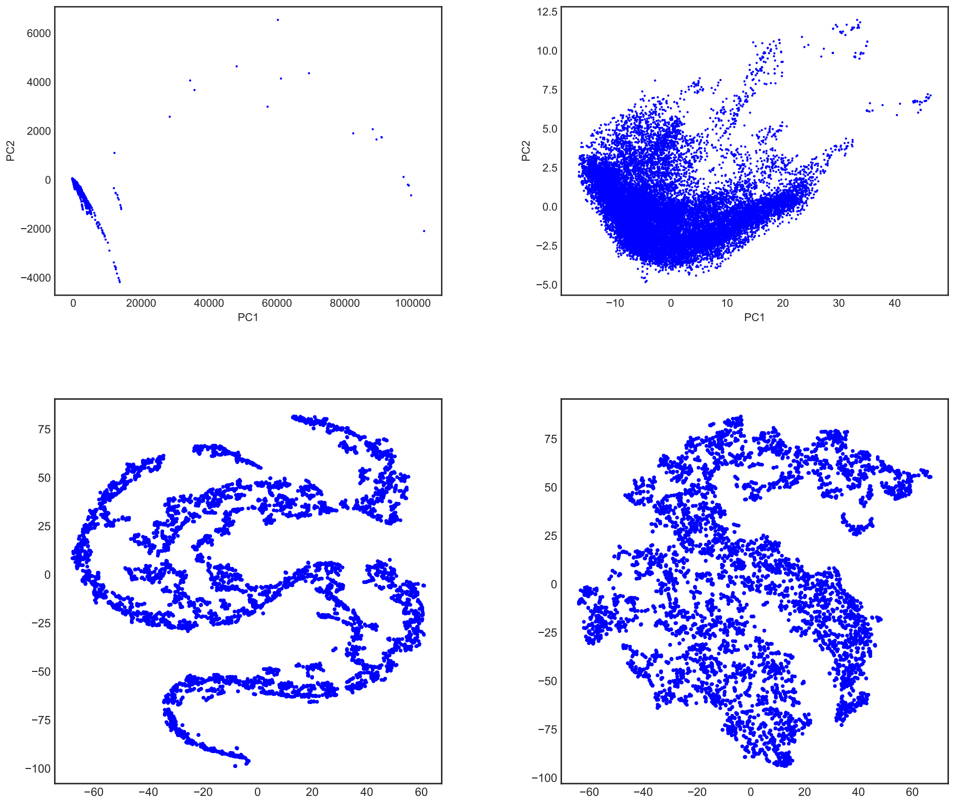


FIGURE 3.9: Visualisation of the Power Spectrum feature applied to a section of Abell2744 of the Frontier Fields. Top left is the feature reduced to projected into 2D using principle component analysis. And the bottom left is a 2D visualisations of the same data using the non-linear technique called t-SNE. We can see there is extensive non linear structure in the data. The PCA plot shows extreme outliers which happen to be the extremely bright areas within the central bulges of galaxies. By applying the natural log we achieve a more standard PCA visualisation at the top right. The bottom right is the same data with the natural log applied visualised using the t-SNE.

3.4.2 Options for representing galaxies

The previous description of image segmentation does not necessarily always identify objects. For example, passive galaxies which have a mostly homogeneous appearance are likely to be segmented in their entirety. However, spiral galaxies contain clear components e.g. spiral arms, that are very distinctive. In addition, segmentation using unsupervised algorithms will not allow us to identify objects, but it will find common areas in images which could also be components of galaxies. Therefore, one cannot just rely on segmentation. In order to group objects it is necessary to apply an unsupervised learning algorithm to cluster a representation of galaxies.

There are two options to represent images of galaxies. The first is to create a fixed size postage stamp, usually a square, around each galaxy. The galaxy maybe enlarged and the depth in the image stretched to create a JPG image (Huertas-Company et al., 2015; Shamir and Wallin, 2014). This is required because computer vision based machine learning systems require a fixed size for each image. However, there are issues with object detection due to contamination. If two or more galaxies are in close proximity the postage stamps will include multiple galaxies which affects the result. In addition, as I am using an unsupervised machine learning process removing extraneous data from the input is important in order for the machine learning algorithms to focus on the data that is most useful in terms of discriminating different types of galaxy.

An alternative method is to create a histogram representation of a galaxy using the segmentation image. This is achieved by identifying which pixels belong to a galaxy in the segmentation and summing all the cluster identifiers. Therefore, the histogram represents counts of the segment types. If the segmentation process is effective then it will identify individual cluster identifiers for sections of spiral arms, for example, are distinct from the central bulges.

A histogram is a useful representation for use in unsupervised algorithms. They can be clustered to find similar histograms (Galaxies), and the can be grouped into their own clusters.

Using this method I must identify the pixels that belong to a galaxy. To do that I use connected-component labelling combined with a multi-level mask. Where the mask represents different threshold intensity levels. The following section describes the connected component labelling algorithm and implementation details.

3.4.2.1 Connected-component labelling

Connected-component labelling is a general term used to describe a process that can identify and label sub-structures within a data set. Each sub-structure is called a component and consists of a set of connected data elements which are considered to be connected if they are joined in some way (for example, vertices that are connected by an edge in an undirected graph). A typical result of the process is a list of uniquely labelled components each consisting of a sub-set of the data elements, where no data element is shared by more than one component. The algorithm is commonly used in image processing to identify and label connected groups of pixels, for example, to identify and extract blobs in binary images. It is not clear when the connected-component labelling concept originated but it has been in use since the 1970s, for example in Hoshen and Kopelman (1976).

Although the general concept is fairly straightforward there are a surprising number of implementation options. Much work has been carried out such as a) the efficient tracing of component outlines or contours (Chang et al., 2004) and b) investigations into algorithm efficiency, considering the relative merits of using a single pass, two pass, or even multiple passes through the data elements (He et al., 2008; Wu et al., 2009). One would expect a single pass algorithm to be the most efficient, however due to the non-sequential memory accesses required by the single pass algorithm the two pass algorithms remain very competitive and execution-time scales linearly with the number of data elements (Wu et al., 2009). Other areas of research into algorithm efficiency include identifying efficient data structures to store and attach labels to data elements such as the 'union-find' data structure (Fiorio and Gustedt, 1996) and research into the parallelisation of various connected-component algorithms including the use of GPUs using NVIDIA's CUDA (Kalentev et al., 2011).

I implement an efficient, sequential version of the algorithm inspired by parts of Wu et al. (2009); Fiorio and Gustedt (1996). However, I deviate from the standard implementations used in image processing by using the algorithm on sub-images (thumbnails) instead of pixel data. Therefore, the term 'data element' in the previous and following sections refers to an individual sub-image. The algorithm proceeds by iterating through the data elements and assigning a label, consisting of an integer value, to each data element. The following steps are performed for each element (the first pass):

1. Retrieve the labels of the neighbouring data elements. Any overlapping or adjacent data elements are neighbours.

2. If there there are no neighbours or none of the neighbouring data elements have labels then create a new label with a unique identifier (an integer that starts with a value zero, incremented for each new label) and apply it to the data element.

3. If any neighbouring data elements have labels then identify the neighbouring label with the smallest unique identifier and assign the label to the data element.

4. Add the unique labels of the neighbouring elements to a list called an equivalence list. This list, containing the integer labels that are considered to be the same, is used at the end of the process to identify all the labels that belong to the same component.

At this point every data element has a label (which may also be shared among many other data elements), each label belongs to an equivalence list and each equivalence list contains all the labels for a unique component.

The second pass is purely a re-labelling process to ensure that every data element in a component has the same label. It proceeds by identifying the equivalent list that the label of each data element belongs to, finds the label in the list with the smallest identifier (the find function of the union-find data structure) and then applies that label to the data element. The output of the algorithm is a list of components and their data elements. The location and size of all the data elements are known and therefore these lists can be used to identify the properties of a component, for example, the width, height and an approximation of its centre.

3.5 Summary

In this chapter I have reviewed some of the options for developing a technique to perform unsupervised analysis of astronomy images. The three questions I raised at the beginning of this chapter were:

1. Which pixel representations should be used?
2. Which unsupervised machine learning algorithms should be used?
3. How should objects be detected and represented?

Through the course of this Chapter I have evaluated the qualities of different pixel representations for their rotation invariance. I have considered features that rely on pixel intensity alone, the spatial frequency of pixel intensity changes, and image gradient based representations. I decided that the Power Spectrum, RIFT and Spin Intensity features should be brought forward and tested for performance.

I evaluated the strengths and weaknesses of various types of unsupervised clustering algorithms that employ different methods such as toplogical mapping (GNG), matrix factorisation (NMF) and density based algorithms (DBSCAN). After applying and evaluating the algorithms on a high dimensional test data set I decided to use the highest performing algorithms of GNG, Hierarchical Clustering and K-Means. These algorithms are used in the experiments described in the following Chapters.

I discussed the object detection and galaxy representation methods available. I decided to use a patch-based histogram representation of galaxies. I chose the Connected-component labelling algorithm in order to identify which pixels belong to galaxies. The output of the Connected-component labelling algorithm can be combined with the clustering of overlapping patches to produce a histogram of each galaxy. This method allows for unusual galaxy shapes and enables galaxies that are in close proximity to processed independently.

In the next chapter I bring together these components and describe the proposed model that incorporates all of the decisions made. In Chapter 5 I will evaluate the hyper-parameters of the model to identify its characteristics and to choose the best features and algorithms for applying the model to the Frontier Fields and the CANDELS datasets.

Chapter 4

The Proposed Model

Section 4.5 of this chapter was previously published in Hocking et al., 2015. Unsupervised Image Analysis and Galaxy Categorisation in Multi-Wavelength Hubble Space Telescope Images, *European Conference on Machine Learning (ECML) Doctoral Consortium*

4.1 Introduction

I now describe the proposed model used in my work. The model is required to perform three steps:

1. Image segmentation
2. Object localisation
3. Clustering morphologically similar galaxies

The ability to segment an image is an important first step as the output of the segmentation process is used to create a representation of individual galaxies in order to categorise them. Therefore, the third step of clustering morphologically similar galaxies can only be performed after the first two steps have been completed.

Although the model has been designed to work on an array of input data the majority of the analyses I have performed use survey images stored in the FITS file format. Galaxy imaging surveys typically provide image files of an area of sky in multiple wavelengths. The model allows multiple FITS files to be used for each area of sky. The first decision when using the model is to select which of the available wavelengths to use in the analysis.

4.2 Image Segmentation

Image segmentation is the automatic process of identifying common areas within images. In order to satisfy two major tasks of segmenting images and categorize morphologically similar galaxies I deviate from the standard method of analysing square images of whole galaxies and instead employ an overlapping-patch based model. By analysing small overlapping patches which are typically much smaller than the sizes of resolved galaxies, the model is able to identify common sub areas and components within galaxies. The segmentation process consists of the following steps:

- Densely extract overlapping fixed size squares of pixel values (image patches) from unlabelled survey images.
- Perform feature extraction by calculating feature descriptors on the overlapping image patches.
- Apply unsupervised machine learning algorithms to create a dictionary of patches using the centres of clusters.
- Create a segmentation image of a survey image using the dictionary of patches.

4.2.1 Patch and Feature Extraction

A fixed sized, sliding window is moved over every pixel in a FITS file. A window size must be chosen by performing tests to identify an optimal size (see Chapter 5). The window contains the intensity values for one ‘patch’. A patch is extracted for every pixel. The pixel intensity values within a patch are then converted into features by applying a feature descriptor for the patch. In the previous Chapter I presented several options of feature descriptor such as RIFT, Spin and the power spectrum.

The patch extraction and feature extraction is calculated for each image file. The choice of image files depends on the selected wavelengths to be included in the analysis. The feature descriptions for each pixel (one from each image file) are concatenated together to form a single vector for the pixel. The concatenation ensures that each vector encodes the information from multiple colours at a single position in the sky.

The vectors for all the pixels form a matrix $n \times m$ when n is the number of patches, and m is the number of pixels in a patch. Each of the m columns (each feature) in the matrix is normalised to have a mean of zero and unit standard deviation.

4.2.2 Application of unsupervised algorithms

Converting all the overlapping patches into features results in a very large data set proportional to the size of the input survey image files. The model uses unsupervised machine learning algorithms to identify the structure of this data, thereby creating a model of this structure. This occurs in two steps. The first is to use GNG to toplogically map the structure of the data with the extracted feature descriptors. A graph or vertices and edges is created which models the latent structure of the data. If the input data is too large, then pixels can be extracted from the images at random.

The graph created by GNG typically has many thousands of vertices. Each vertice can be viewed as a position in Cartesian space and models any data point (image patch) in close proximity, where proximity is measured using the Euclidean distance between the vertex and the patch. Another way of perceiving the graph produced by GNG is as a method to sort the number of patches into groups, where each vertex represents a bin of similar patches. In this way I create a dictionary of patches, where each vertex represents a visual item in the dictionary.

The graph typically contains thousands of vertices. In order to identify areas of interest the graph is cut using agglomerative hierarchical clustering.

Both the graph and groups identified by the agglomerative clustering can be thought of as a dictionary of patches. The graph represents a more detailed dictionary containing thousands of patch types, and the agglomerative groups a coarser grained dictionary containing hundreds or up to a thousand patch types.

4.2.3 Segmentation of survey images

The first steps are to select a survey image and then follow the overlapping patch and feature extraction process: overlapping patches are extracted using the sliding window technique, feature descriptors are calculated for each patch and the data normalised as shown in Section 4.2.1. However, instead of then applying the unsupervised machine learning algorithms, I use the dictionary of patch types instead. If the GNG/HC graph successfully maps the latent structure of the data then the graph can be used as a model by identifying which vertex in the graph is most similar to a feature descriptor. Each vertex represents a patch type. So by identifying the vertex most similar to a feature descriptor I identify the patch type used as an identifier to tag the corresponding pixel. This process is repeated by tagging all the pixels in an image with a patch type identifier.

Normally I would consider this identifier as a pixel's classification. However, the word 'classification' in the machine learning community strongly implies a supervised process, which we do not use and so I use the term 'identifier' instead.

The result is a new image of the same size as the original with an identifier placed at each pixel position. A segmentation image is then produced by selecting a colour for each identifier and producing an image with all identifiers replaced with their corresponding colours.

4.3 Object Localisation

The data consists of large images containing 1000s of objects which we must locate. The segmentation process as described has no notion of an object. It identifies similar patches typically smaller than galaxies. In order to identify a galaxy the model uses a two stage process to detect objects: identify the background level above which a pixel is considered to be the signal from a galaxy and use a connected component labelling algorithm to perform simple blob detection, where each blob is considered to be a galaxy.

4.3.1 Identifying the Background

I identify general object positions by removing the background which is identified as pixel values below a threshold level. The threshold is automatically calculated using a method called sigma clipping (using the Python astropy library (Robitaille et al., 2013; The Astropy Collaboration et al., 2018)). Sigma clipping involves interating over the pixel intensity values of an image or representative image subsection. Each iteration calculates the median pixel value and then rejects pixels that are more than a specified number of standard deviations away. The rejected pixels have higher intensity typically because they represent detections of light from galaxies. The threshold is finally identified when no more pixels are rejected or a specified number of iterations has occurred. The final threshold value is the standard deviation from the last iteration.

When applying this technique to crowded fields I use a multi-level threshold. For example, the Frontier Fields consist of crowded and very bright galaxies within the cluster. The galaxies in the cluster have a very bright extended emission over large areas. The background level around these galaxies can be considered to be higher. By using multiple values for the initial standard deviation, the technique can identify individual galaxies in the central parts of images of galaxy clusters.

4.3.2 Galaxy Localisation

The next step is to create a notion of a galaxy that can be used with unsupervised machine learning algorithms. By using connected-component labelling, described in Chapter 3, in combination with the identified background level(s) it is possible to identify which pixels belong to or are in close proximity to a galaxy. The connected-component labelling algorithm identifies blobs (sections of overlapping patches) in the survey image above the threshold pixel level. It labels each blob with a unique identifier - a galaxy identifier. In addition, once I know which pixels belong to a galaxy I can use simple statistics to identify aspects of each galaxy such as the approximate centre, shape and size. For example, by averaging the locations of all the pixels of a galaxy I identify its approximate centre.

The output of the background level detection and galaxy localisation is a list of galaxies, their locations and shape attributes. However, I can improve this categorisation by grouping galaxies using the data acquired in the segmentation process.

4.4 Clustering morphologically similar galaxies

To use unsupervised machine learning algorithms to further categorise galaxies a vector representation of each galaxy must be created. This model uses a histogram as this representation for each galaxy. The previous step identifies which pixels belong to a galaxy and each of those pixels has an identifier allocated by the segmentation process. Each bin in the histogram is an identifier from the global set. The total number of bins in the histogram corresponds to the size of the dictionary. The pixel identifiers are counted within each galaxy and the corresponding bins are set in the histogram. The result is a sparse vector with positive values in the bins representing the identifiers that occur in the galaxy. This process is repeated for each galaxy identified by the previous galaxy localisation step.

The histogram vectors representing all the galaxies are combined to form an $m \times n$ matrix, where m is a galaxy vector and n is the total number of bins. The matrix of galaxies is then weighted using Term-Frequency Inverse- Document-Frequency (Sparck Jones, 1972, TF-IDF). This is important as it not only down weights the identifiers of the patch types that occur most frequently across all of the galaxies, but also creates an element of scale invariance by normalising the counts for different sizes of galaxy as larger galaxies consist of many more pixels. Groups of similar galaxies are then identified by clustering matrix using K-Means or agglomerative clustering.

4.5 Applying the Clustering Model to Segment Images

I now apply the first part (image segmentation) of the model to segment images of the Frontier Fields. These are the first results of applying the technique. The best model hyper-parameters are not clear at this point. The next chapter describes the tests of the whole model that I carried out to identify the best hyper-parameter values and algorithm options. Section 5.6.2 describes the considerations when choosing the patch size.

4.5.1 Segmenting Frontier Field Images

I select the Frontier Fields image Abell 2744 for segmentation. The overlapping patches are extracted and feature descriptors calculated for each patch. The matrix is normalised and then GNG is applied. The resulting GNG graph contained over 7,000 vertices. This number is too

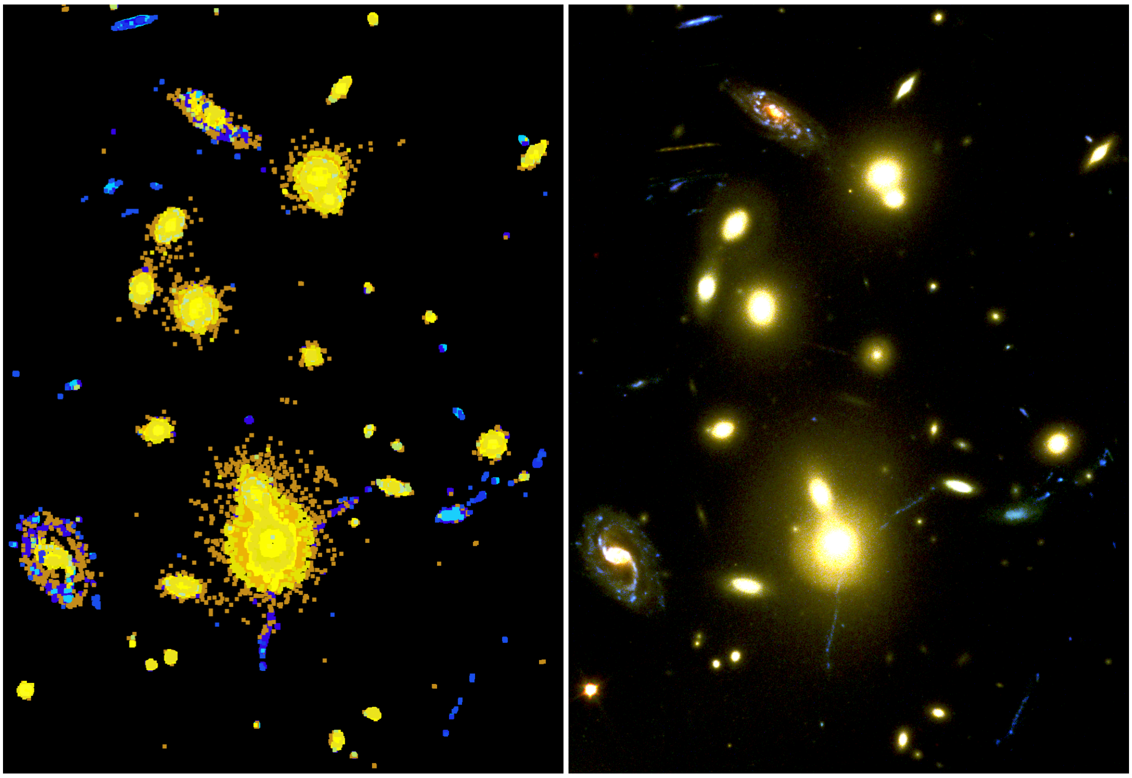


FIGURE 4.1: A sub section of the image representing the model (left) and the thresholded HST image (right) of the galaxy cluster Abell 2744. Blue colours in the processed image highlight the unsupervised clusters that represent star-forming regions in the spiral and lensed galaxies. Yellow colours correspond to the unsupervised clusters that represent passive elliptical galaxies and the central passive regions of spiral galaxies.

great to easily visualise to understand the latent structure of the pixel data that the graph represents. I therefore used agglomerative clustering to cut the graph into a more manageable number of clusters. This produced a tree structure that represents a hierarchy of merged clusters. Each node in the tree structure represents a new cluster consisting of the two hierarchical clusters with the greatest similarity. Cluster similarity was measured using average linkage and the Pearson correlation distance with an additional penalty. The Pearson correlation distance measures the dissimilarity between the centroids of two GNG clusters. If the centroids are equivalent over the majority of the fifteen sample values then the distance is small.

The recursive clustering process was continued until all clusters had merged. A top down search was performed on the tree structure to identify the hierarchical clusters with a similarity greater than a threshold value. This resulted in 253 clusters with the largest 40 clusters representing over 97% of the samples.

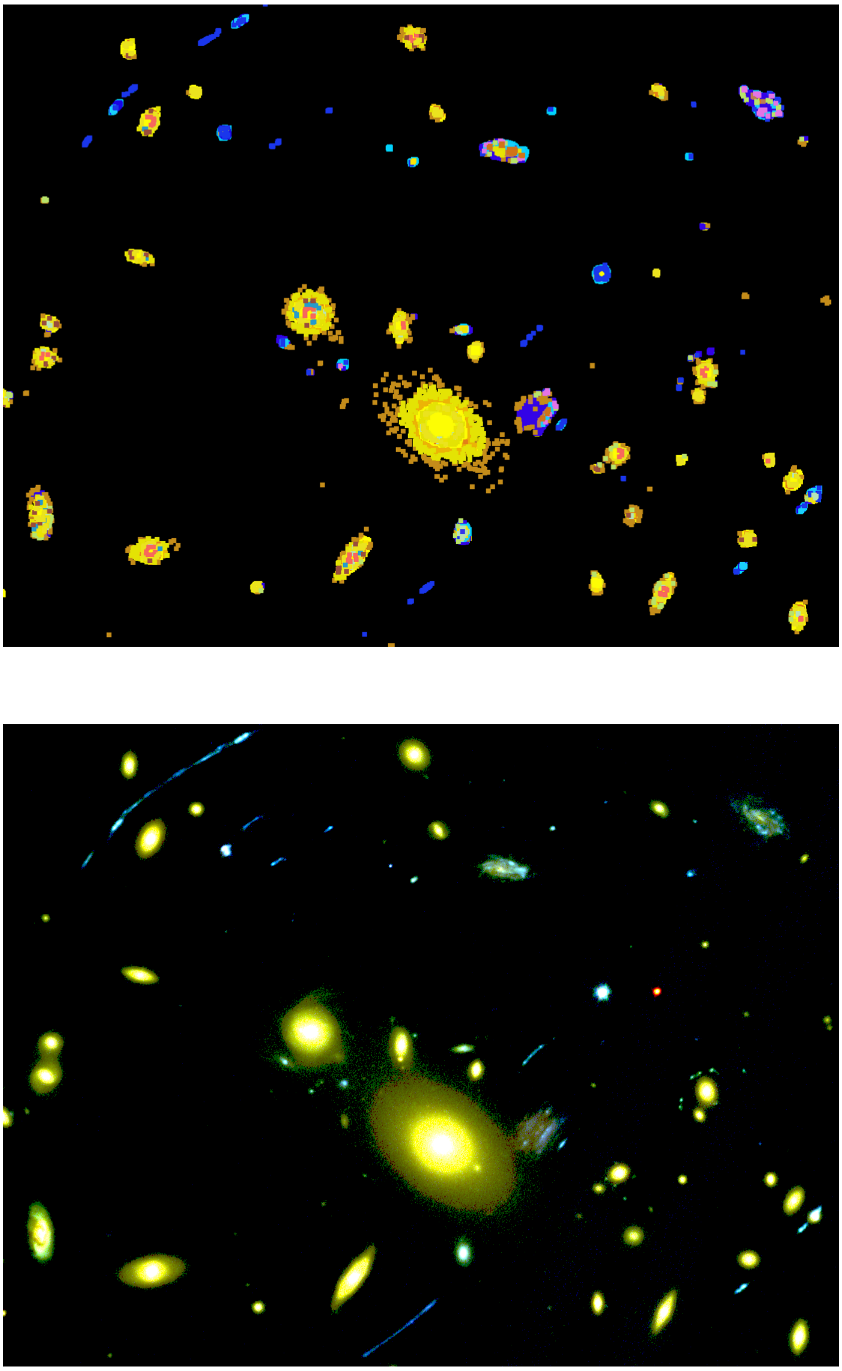


FIGURE 4.2: The processed image at the top displays the result of applying the model to new HST image of galaxy cluster MACS0416. Processed image of the MACSJ0416 galaxy cluster. HST composite RGB image of the MACSJ0416 galaxy cluster.

4.5.2 Post Processing

In order to analyse the clusters I started with a blank image and added colour to the pixels (a different colour corresponding to each cluster) at the original positions of the samples. This image was then compared to a false colour RGB image, which was created by combining the original red, blue and green HST thresholded images. This confirmed that the clusters identified by the machine learning process correspond to the distinct star-forming and passive galaxies in the RGB image, as illustrated in Fig 4.1.

It is likely that the clusters are heavily influenced by colour. This is because a patch from each of the three colour images is concatenated to form a single vector therefore the relative colour differences across the three colours are encoded in the vector. These are then clustered and if the relative colour differences are large then this could dominate. The power spectrum identifies changes in spatial frequency therefore any steep gradient changes will be reflected in the power spectrum. It could be argued that the small patch size means that morphology may not have a strong influence, as the power spectrum may only identify the 'smoothness' of the individual patches. This is likely to be the case in Fig 4.1, however when we combine the power spectrum of the overlapping patches in a whole galaxy, this is capturing the changes in smoothness across the galaxy and therefore morphology is likely to provide a stronger contribution to the clustering.

4.5.3 Preliminary Results

The model created by applying the unsupervised learning steps in Section 4.5 to Abell 2744 was tested for its ability to generalise to other galaxy clusters, using MACSJ0416 as an example. The model created in Section 4.5 was then applied to the new feature matrix by using a nearest neighbour calculation with the Euclidean distance metric to identify the nearest cluster to each sample. The results were used to create a processed image of clusters for the MACSJ0416 galaxy cluster using the same process as in Section 4.5.2. Fig 4.2. displays the results. The processed image was accurate and correctly categorised each overlapping patch with the same colour (or cluster) that appeared in the processed image for Abell2744. Each star-forming and passive galaxy in the processed image shows the identified clusters representing the star-forming and passive galaxies in the HST image.

4.6 Summary

I have established a model that can segment survey images, locate objects and group similar objects. I have presented initial results of segmenting Frontier Field survey images but I have yet to establish the correct model hyper-parameters of algorithm choices that have the highest performance. The next chapter describes extensive hyper-parameter testing to identify the optimum configuration to perform clustering of morphologically similar galaxies in *HST* data. I then use the optimum configuration to analyse *HST* CANDELS survey in Chapter 6.

Chapter 5

Model Selection

Sections 5.4, 5.5 and 5.6 were previously published in Hocking et al., 2017 Mining Hubble Space Telescope Images, *International Joint Conference on Neural Networks (IJCNN)*.

5.1 Introduction

I have established an unsupervised machine learning model to analyse survey images, and I have initially applied it to segment Frontier Fields images. However, there are a number of hyper-parameters, such as patch size and GNG graph size, that need to be chosen in order to produce the most effective galaxy categorisation. Unusually for an unsupervised machine learning model I adopt a validation and test data split to perform the optimum hyper-parameter search. Performing the tests in this way enables me to evaluate how well the model can generalise to new images that are not part of the initial analysis. In other words I am testing whether the model created by analysing the latent structure of a survey image is also applicable to new images. I expect this to be true for data that has the same characteristics such as survey data observed using an individual telescope and to the same depth. The following Chapter describes the results of performing many thousands of tests to cluster galaxies to identify the combination of parameters that are most effective. But first of all I evaluate the technique's performance in detecting objects in survey images.

5.2 Localisation Performance

The method of identifying objects in survey images used in this technique is substantially different from existing techniques such as those used by Source Extractor (Bertin, 1996). I now consider the relative performance of the source identification process by comparing the catalogues produced by the connected component-labelling technique to official catalogues provided for the Frontier Fields and CANDELS.

5.2.1 Frontier Fields

In contrast to most optical imaging surveys, such as CANDELS, the Frontier Fields present an interesting problem from the point of view of source detection. The images of the strong lensing clusters contain very bright, concentrated and extended elliptical galaxies. The background level surrounding and between the elliptical galaxies is very high. This causes a distinct problem for source identification as the background level is often higher than galaxies that appear further away from the central cluster.

I evaluated the localisation performance of the MACSJ0416 Frontier Field by comparing the catalogue produced by the machine learning technique with the Merlin catalogue (Merlin et al., 2016)[1] which was produced using software from the ASTRODEEP project[2]. Merlin performed their source detection using the *HST* F160W images only. They employed an extensive process to account for the extended light surrounding the brightest galaxies of the cluster. This is unusual as most surveys do not contain such extended sources, something Merlin et al. (2016) describe as 'intra-cluster light (ICL)'. To improve source detection they removed the ICL by employing a complex five-step 'cleaning' approach, incorporating masking, GALFIT (Häußler et al., 2013) and GALAPAGOS (Barden et al., 2012) to perform profile fits and filtering to remove the light from the foreground bright sources. With the ICL light removed Source Extractor was used in two configurations: the first to detect the bright sources, and the second to detect the fainter sources. They found that if they did not use the five-stage approach Source Extractor identified many false detections due to the ICL.

[1] Frontier Field catalogues and processed images can be downloaded from http://www.astrodeep.eu/frontier-fields/
[2] http://astrodeep.eu

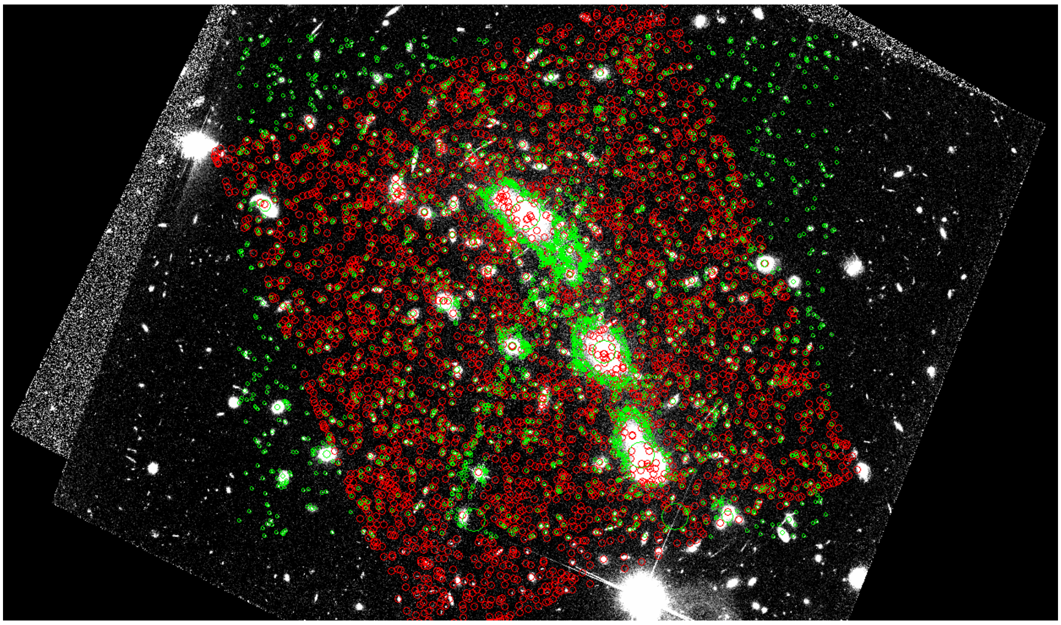


FIGURE 5.1: This shows the *HST* image of MACSJ0416 in the F606W filter. The red circles represent detections listed in the official catalogue Merlin et al. (2016) and the green circles are the sources listed in the machine learning catalogue.

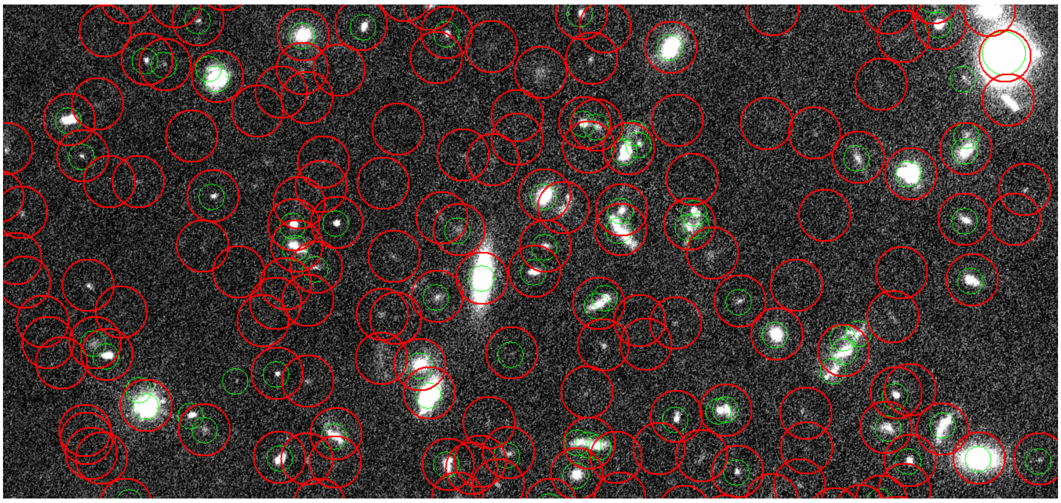


FIGURE 5.2: A detailed view of the sources identified in Frontier Field MACSJ0416. This image is the F606W image from the *HST*. The red circles are the sources listed in the Merlin et al. (2016) catalogue and the green circles are the sources listed in the machine learning catalogue. This image is centred at 64.04925 (RA) and -24.0807 (Dec.) and is 40″ × 19″.

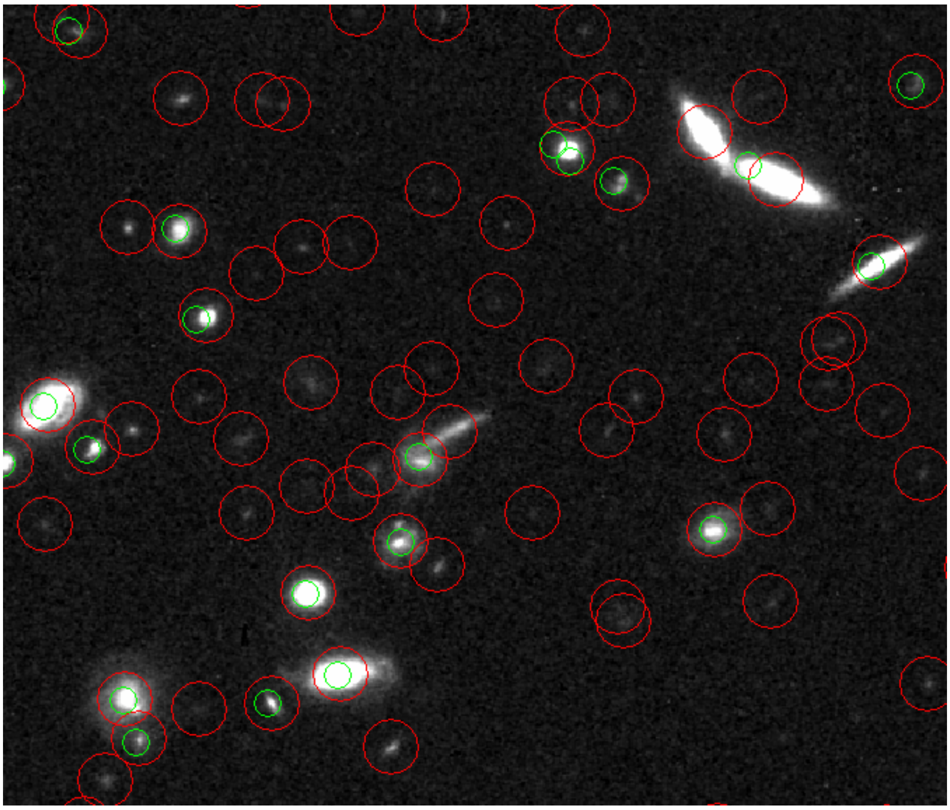


FIGURE 5.3: A detailed view of the sources identified in part of the COSMOS CANDELS field. This is the F160W image from the *HST*. The red circles are the sources identified in the Skelton et al. (2014) catalogues and the green circles are the sources identified in the machine learning catalogue. This image is centred at 150.0971 (RA) and 2.31093 (Dec.), it is $34'' \times 28''$.

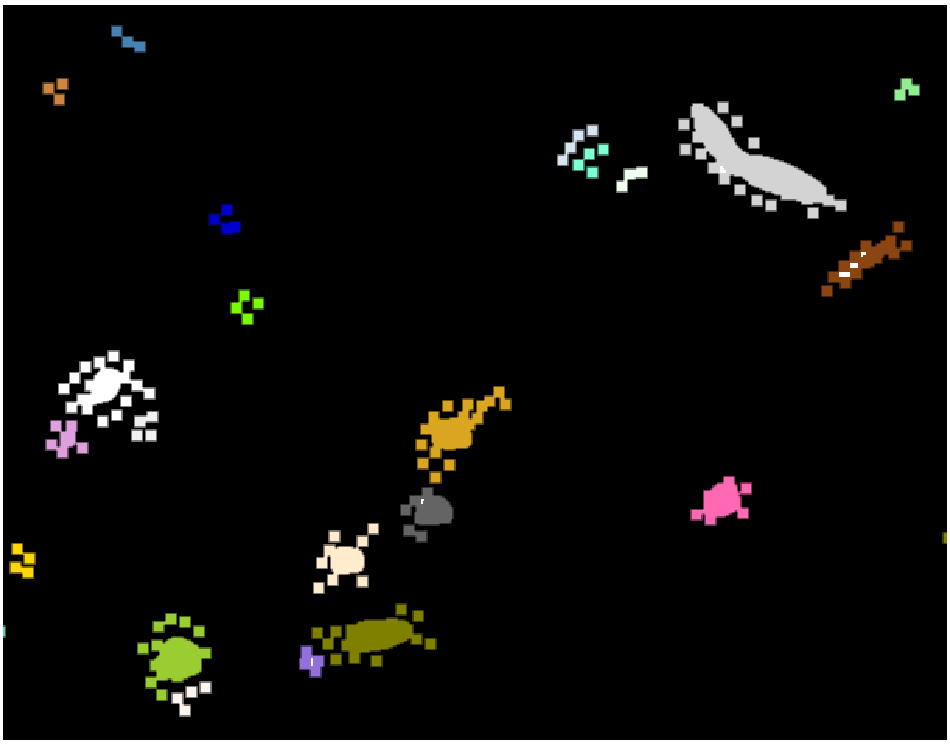


FIGURE 5.4: This is the segmentation map produced by the machine learning algorithm of the same area as Figure 5.3. A group of pixels with the same colour marks the identification of an object. This shows the pixels considered to be single objects by the machine learning technique.

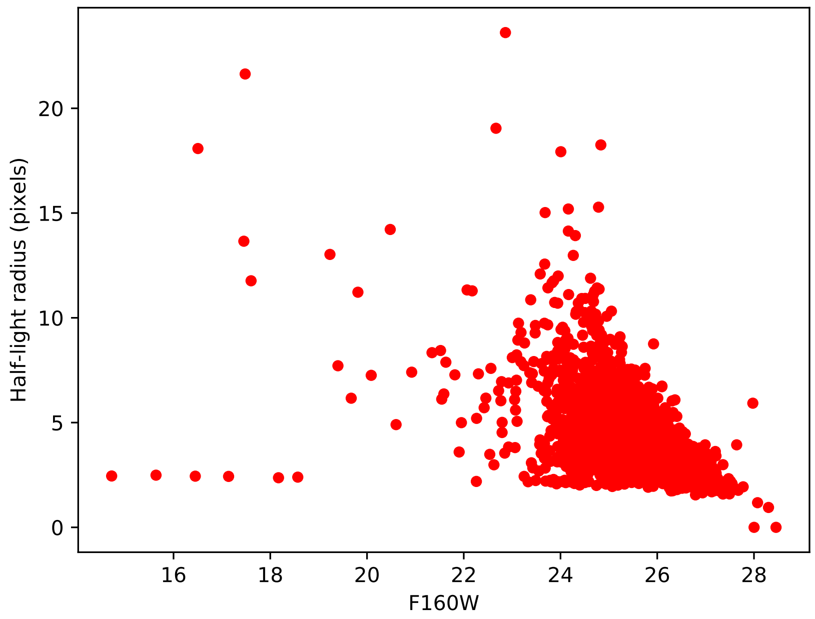


FIGURE 5.5: This shows the COSMOS galaxies from the official CANDELS catalogue that did not appear in the machine learning catalogue. The half-light radius is the FLUX_RADIUS parameter from Source Extractor. It is the circular aperture radius enclosing half the total flux of the galaxy. The vast majority of the galaxies have a magnitude of 24 or higher suggesting that the threshold level used for source detection was too high for the machine learning technique to identify these objects.

The machine learning technique's source detection is performed on a combination of F435W, F606W and F814W imaging. I calculated the one sigma background level for each filter using sigma clipping and then obtained lower and higher threshold values by multiplying each background level by 5 or 10 respectively (described in Section 4.3.1). I first applied the higher threshold to each image to detect the brightest sources, which would otherwise be lost within large areas of fainter emission if only a low threshold were used. I then applied the lower threshold to each image to detect fainter sources. When a bright source lies within a faint source the faint detection is ignored (see Sections 3.4.2, 3.4.2.1 and 6.2.1.3 for more details). The use of multiple threshold levels was adopted to enable the separation of the brighter central cluster elliptical galaxies. Figure 5.1 shows the results, while it is clear that the galaxies in the central bulge are detected, there are also many false detections within these large galaxies at the point where the brightness drops below the higher threshold level. The false detections in the ICL do not affect the results in this Chapter because I compare the clustering results of 120 galaxies with known classifications.

I used the ‘match_to_catalog_sky’ function in Astropy’s coordinates package (The Astropy Collaboration et al., 2018) to match galaxies between the two catalogues. A source separation threshold of $0.25''$ was used for matching. There were 3058 galaxies from the official catalogue in the overlapping region. These were matched to 1310 detections in the machine learning catalogue representing a match of 42%. The difference between the two catalogues is illustrated in Figure 5.2. This figure presents a region outside of the central cluster and shows that the objects not present in the machine learning catalogue are the smaller, fainter galaxies. This indicates that the background level used for source detection was too high. It is also not clear whether there is a difference in source detection due to the use of the shorter wavelengths F435W, F606W and F814W by the machine learning technique in contrast to the use of the longer wavelength F160W used by the official catalogue.

5.2.2 CANDELS

The CANDELS survey is more straightforward than the Frontier Fields survey, from a source detection point of view, as the images do not contain such detailed imaging of strong lensing galaxy clusters. I matched the CANDELS catalogue produced using F814W and F160W images with the 3D-*HST* catalogues from Skelton et al. (2014) who also used F160W for source detection. For the purposes of evaluating localisation performance I restricted my analysis to a section of the COSMOS field.

The background level was detected using the same approach of using sigma clipping. A single threshold was calculated for the whole of COSMOS for each of F814W and F160W images. The catalogues were matched using Astropy and of the 5136 galaxies in the overlapping area of Skelton et al. (2014), 1427 were matched to the machine learning catalogue, giving a 27.7% match. Figure 5.3 shows a detailed image comparing the sources from Skelton et al. (2014), in red, with the sources from the machine learning catalogue, in green. Once again we see that the objects with lower apparent magnitude are not being identified, most likely due to the background level being set too high.

Figure 5.4 reveals the areas of the image that are identified as objects by the connected-component labelling process. It corresponds to the image seen in Figure 5.3. Within the image we can see that overlapping galaxies are considered to be single objects. This has the advantage that interacting galaxies will be identified as single objects and grouped accordingly. The disadvantage is that chance alignments will also be considered as single objects. The Skelton catalogue identifies these objects individually. The Figure also shows that the faint, smaller objects are too faint for the threshold that was set.

Figure 5.5 shows the magnitudes and flux radius of the galaxies that were missing from the machine learning catalogue. The vast majority of the missing galaxies had a magnitude greater than approximately 24 ABmag indicating that the cause of the missing galaxies is the background threshold level.

Overall, the primary differences between the catalogues for COSMOS occur due to the machine learning technique using a single value for the background level over the whole field and the this level being too high and therefore excluding fainter objects. A variable background threshold mask over the entire field would improve the source detection of fainter objects.

I now evaluate the technique's performance in grouping galaxies of known types and identify the optimum hyper-parameters and model configuration.

5.3 Data

I use the FF dataset described in Section 2.4 of Chapter 2. The FF images contain good examples of five types of galaxy including lensed galaxies and galaxies exhibiting strong lensing effects. Examples of each type from the *HST* data can be seen in Figure 5.6.

In my experiments I use data from three filters and three areas of sky. There are a total of nine FITS files, each approximately 500Mb in size, so in total 1.5GB of data for each area of sky.

I create unsupervised models using one area of sky (sky area 0). Each model uses different preset parameters. I then validate the models on the second area of sky (sky area 1) by comparing the model results with classifications provided by an independent astrophysicist (N.K. Hine, private communication). Finally to test for generalisation I evaluate the clusters produced using the best unsupervised model on the third area of sky (sky area 2). More details on cluster evaluation can be found in the next section.

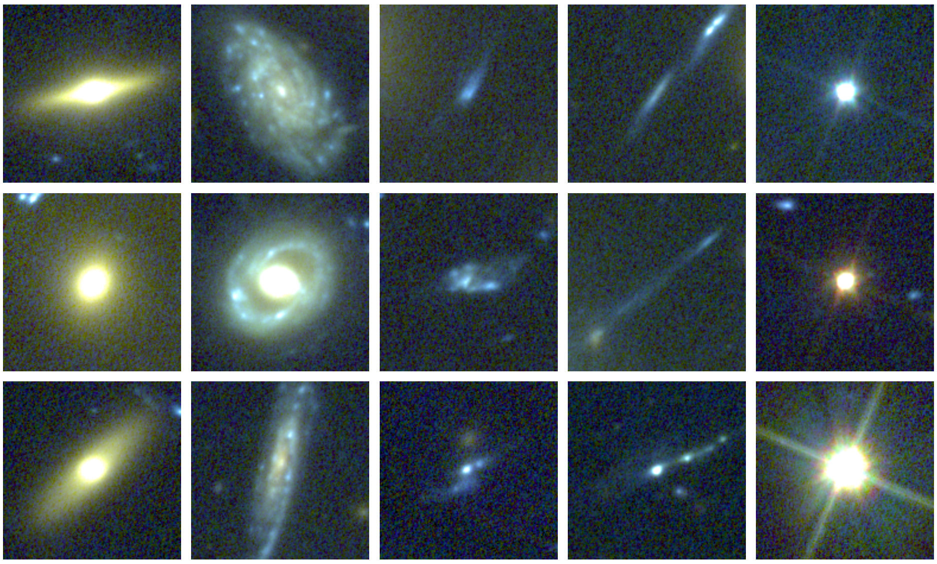


FIGURE 5.6: Galaxies and point sources from the *Hubble Space Telescope* Frontier Fields survey. Each column contains three examples of a object type starting on the left with elliptical/lenticulars, then spirals, background star-forming galaxies, lensed galaxies and finally point sources. The RGB channels are F435W, F606W and F814W. Images produced using data from: NASA and STScI

5.4 Performance Evaluation

In order to evaluate the models I compare the results of analysing galaxies in sky area 1 with human acquired classifications. The technique starts by extracting features for each overlapping patch from sky area 1 and identify the most similar dictionary example. I then combine with the information from the Connected component labelling algorithm to create galaxy vectors. I now compare the galaxy vectors to the galaxy type centroids from the model construction (the output from K-means). Each galaxy vector is classified based on the most similar galaxy centroid.

To evaluate clustering performance I compare the clusters identified in sky area 1 with the classifications provided by an independent astrophysicist (see Table 5.1). For the purposes of cluster evaluation each galaxy type is considered to be a cluster. I used the following evaluation measures:

1. Adjusted Rand Index (ARI) (Vinh et al., 2009)

   This is a similarity score between two clusterings with an adjustment for the chance grouping of elements. A score of 1 indicates a perfect match and a low score near 0 indicates a very poor match. I treat the human classified labels as the first clustering and the cluster identifiers provided by the technique as the second clustering.

2. The confusion matrix

   I use the confusion matrix to provide a more detailed analysis of the clustering results. I use both *precision*[3] and *recall*[4] (Raghavan et al., 1989). The confusion matrix can be seen in Table 5.2. In order to produce this confusion matrix I identified which clusters produced by the machine learning system correspond to a particular morphological galaxy type. This is done by manually comparing the clusters assigned to the 120 galaxies (by the model) with the classifications provided by the astrophysicist. For example, the astrophysicist has classified 40 galaxies as elliptical and the machine learning system has put 35 of those galaxies into distinct clusters (along with 5 false positives), therefore these clusters are identified as elliptical.

Once the optimum pre-set parameters have been identified using the the sky area 1 dataset, I do the same procedure on sky area 2.

5.5 Experiments On Identifying Types of Galaxies

The main aim of this Chapter is to validate whether the proposed unsupervised machine learning models can identify different types of galaxies, especially interesting and difficult types such as lensed galaxies. In these experiments I have created the models using the model construction dataset with different sets of pre-set parameters. I then validated the models on the sky area 1 dataset to identify the most suitable parameters. The results shown are obtained from applying the model to the sky area 2 data set.

5.5.1 Identifying Types of Galaxy

The source data contains five main types of galaxies. Examples of each type are shown in Figure 5.6. The purpose of this experiment is to identify how effective the technique is at producing clusters that represent these types. Table 5.2 shows the confusion matrix for the results of applying the pixel intensity power spectrum feature to this problem. What I see is the technique is successful at identifying elliptical galaxies and spiral galaxies. However, the majority of the galaxies showing large strong lensing features are in the same clusters as the background

[3] Precision $P = TP/(TP+FP)$ where TP is true positives and FP is false positives.

[4] Recall $R = TP/(TP+FN)$ where TP is true positives and FN is false negatives.

TABLE 5.1: The list of galaxies classified by the astrophysicist in two sky areas. These are used to evaluate the clusters produced by the unsupervised model and to test for generalisation.

	Elliptical	Spiral	Background	Lens	Point Source	Total
Sky area 1 (hyper-parameter search)	51	19	27	10	13	120
Sky area 2 (test for generalisation)	40	17	41	19	3	120

TABLE 5.2: The confusion matrix to show the effectiveness of categorising types of galaxy in sky area 2. The clusters identified by the unsupervised method are compared to the classifications provided by an astrophysicist. This table shows the results for the pixel intensity power spectrum vector representation. We can see that the clusters effectively categorise spiral and elliptical galaxies, the lensed galaxies were clustered together with background galaxies. However, RIFT and Spin Intensity feature representations were more successful at distinguishing background from lensed galaxies.

		Results from Clusters					
		Lens	Elliptical	Spiral	Background	Point Source	Total
Estimate from Astro-physicist	Lens	5	0	1	13	0	19
	Elliptical	1	35	3	1	0	40
	Spiral	0	3	14	0	0	17
	Background	2	1	3	35	0	41
	Point Source	0	1	1	0	1	3
	Total	8	40	22	49	1	120

galaxies. It should be noted that the lensed galaxies are effectively as subset of background galaxies which I chose to separate out into a separate group as they are of particular interest to astrophysicists. It is therefore not surprising that there is some confusion between these two classifications. However, in the next section I identify a configuration that was more successful at distinguishing lensed galaxies.

5.5.2 Detecting Strong-lensing features

Table 5.3 shows the results of applying the model to sky area 2. We see that pixel intensity power spectrum has the best recall however the precision of 0.39 is confirmation that pixel intensity power spectrum is less successful at distinguishing lensed galaxies from background galaxies. RIFT and Spin Intensity features have fewer misclassifications of background galaxies but are more prone to false negatives. Therefore, although the RIFT and Spin Intensity features result may miss some lensed galaxies the cluster produced by the model will contain a high proportion of lensed galaxies.

TABLE 5.3: The results of identifying lensed galaxies in sky area 2 using three different feature representations. Rotation invariant feature transform was the most effective.

Feature	ARI	Precision	Recall
RIFT	**0.48**	**0.69**	**0.58**
Spin Int.	0.44	0.67	0.53
PS	0.24	0.39	0.74

5.6 Discussions on Pre-set Parameters

I perform extensive experiments to identify the optimum pre-set parameter combinations and algorithm options. I used a model construction dataset (sky area 0), a dataset for validating parameter combinations (sky area 1) and a dataset to test for generalisation (sky area 2) described in Section 5.3. Models were created for each pre-set parameter combination. The models were evaluated on the sky area 1 data set.

This is an unusual method to evaluate an unsupervised machine learning model, because it is a test of generalisation. Most unsupervised machine learning techniques are used for data mining, to search an entire dataset. However, I am developing a model that learns structure in one dataset and identifies similar structure in a second. I therefore need to test that the model can generalise to new data.

The following pre-set parameters were evaluated:

1. Pixel intensity power spectrum (PS), RIFT & Spin Intensity descriptors.
2. Patch size, number of radial profile bins and bin size.
3. GNG: graph size of 5000, 15000, 25000 graph nodes.

TABLE 5.4: The following preset-parameter options provided the best adjusted rand index scores for the application of an early type (elliptical/lenticular) and late type (spirals/background/lensed) galaxy cut for each of the three chosen features. These results used the Pearson Correlation encoding, agglomerative clustering and 5000 GNG nodes to categorise the galaxies. The patch sizes were 8px × 8px for PS, 16px × 16px for Spin Intensity and 32px × 32px for RIFT.

Feature Descriptor	ARI
PS	**0.84**
Spin Intensity	0.64
RIFT	0.55

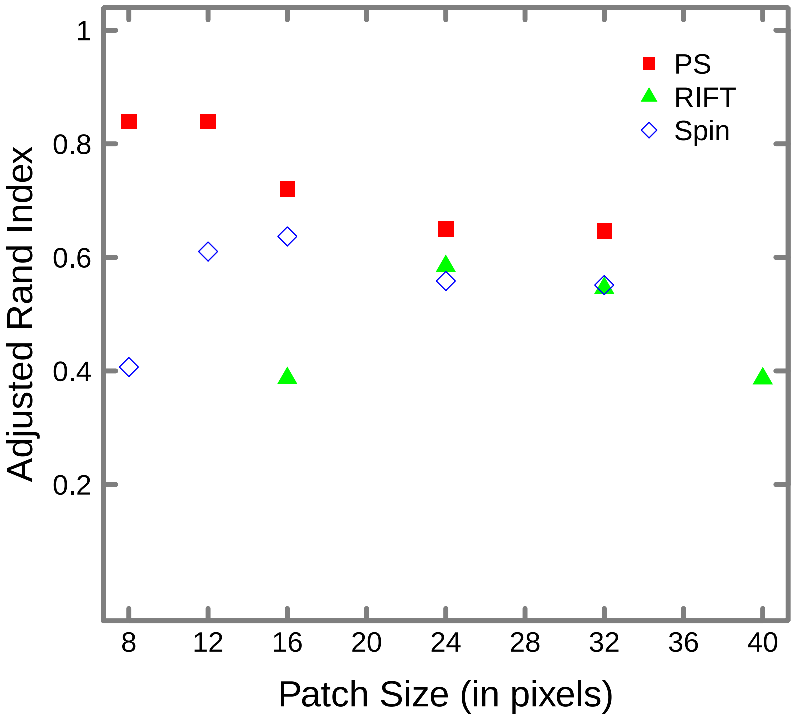


FIGURE 5.7: The results from the applying the technique to sky area 1 are presented for different patch sizes of each feature for the purpose of late type and early type cut. 8 and 12 pixels square patches are the optimum for the pixel intensity power spectrum representation. The pixel intensity power spectrum and Spin Intensity features were tested on patches ranging from 8px × 8px to 32px × 32px patch sizes, RIFT was tested with patch sizes from 16px × 16px to 40px × 40px

4. Galaxy encoding similarity measure: Pearson correlation coefficient or Euclidean distance.

5. Clustering algorithm used to cluster galaxy vectors: K-means or agglomerative clustering.

5.6.1 Effect of Representation Type

I have chosen three representations: pixel intensity power spectrum, RIFT and Spin Intensity. Table 5.4 shows the best results for the application of late type (spiral/background/lensed galaxies) and early type (elliptical/lenticular) galaxy cut. It is clear that for this application the pixel intensity power spectrum shows a significantly higher performance than the others. This suggests that pixel intensity power spectrum is most effective at encoding relative colour differences of the source data.

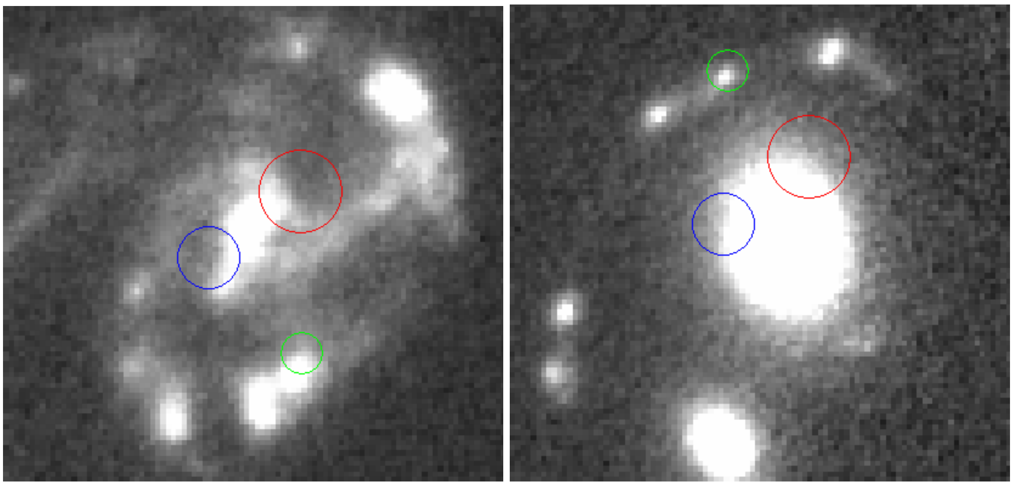


FIGURE 5.8: Two galaxies in the Frontier Field MACS0416. The green circles have diameters of 8px (0.24″), the blue circles have diameters of 12px (0.36″) and the red circles have diameters of 16px (0.48″). These circles illustrate the sizes of overlapping patches. The images are 12.36″ by 11.79″.

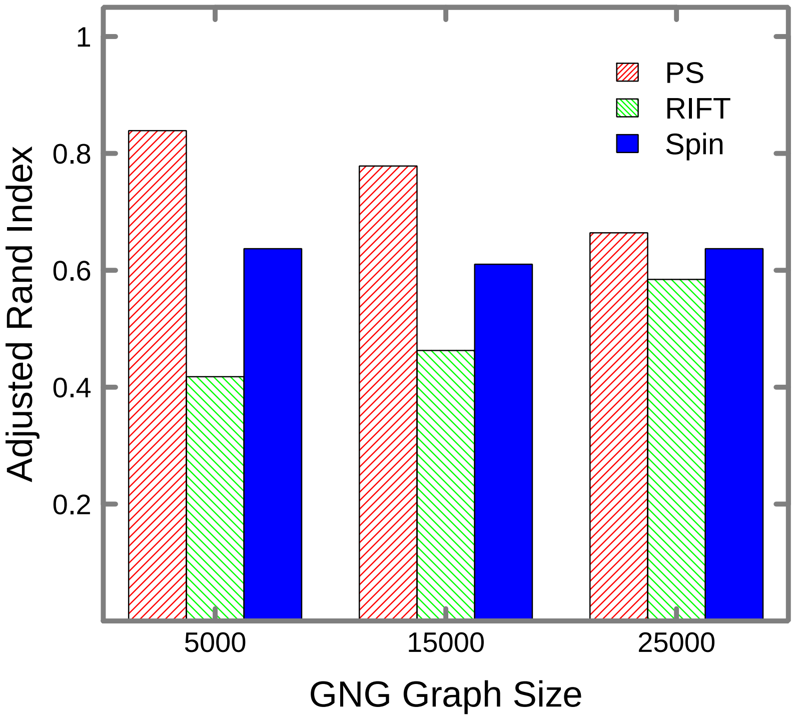


FIGURE 5.9: A chart of the effect of the number of GNG nodes, the GNG graph size, for each of the representation types. Three different graph sizes were validated: 5000, 15000 and 25000 nodes. Pixel intensity power spectrum can produce good results even with a smaller GNG graph, however, the RIFT representation requires a larger graph size.

5.6.2 Effect of Patch Size

The patch size refers to the size of the overlapping square sub images that are extracted from the source data. When choosing the patch size, there is a trade-off between the size of patches and the computational expense. However, larger patches also contain more morphology. In my study I have found that, for the Frontier Fields data, patches smaller than size than 8px by 8px ($0.24''$ by $0.24''$) contain little morphology. Therefore, I consider a range of sizes starting from 8px by 8px.

Each pixel in the source data is $0.03''$. Figure 5.7 shows that the peak performance of the pixel intensity power spectrum occurs with smaller patch sizes of 8px by 8px ($0.24''$ by $0.24''$) and 12px by 12px ($0.36''$ by $0.36''$) and then declines with larger patch sizes. Both Spin intensity and RIFT have similar performance profile as the patch size is increased with performance peaking at 16px by 16px ($0.48''$ by $0.48''$) for Spin Intensity and 24px by 24px ($0.72''$ by $0.72''$) for RIFT. Figure 5.8 shows the relative sizes of three different circles representing patches with sizes of 8px (green), 12 pixels (blue) and 16 pixels (red) compared to typical galaxies in the images.

5.6.3 Effect of GNG Graph Size

The GNG would be expected to produce a more accurate estimation of data density as the size of the graph increased. However, we see in Figure 5.9 is that the pixel intensity power spectrum feature achieves better results with a smaller GNG graph, whereas the RIFT feature requires a much larger graph size, and the performance of the Spin Intensity feature does not vary significantly with different graph size.

5.6.4 Effect of Encoding Similarity Measure

I tested the effect of using two distance measures to calculate similarity: the Euclidean distance metric and the Pearson correlation coefficient. Figure 5.10 shows that the Pearson correlation coefficient comprehensively outperforms Euclidean distance. It achieved the best ARI scores when using the pixel intensity power spectrum, RIFT and Spin Intensity. The top 10 results for all parameter combinations when using pixel intensity power spectrum and RIFT did not include any result that used the Euclidean distance. For the Spin intensity feature only the 7th and 10th positions in the top 10 best results used Euclidean distance.

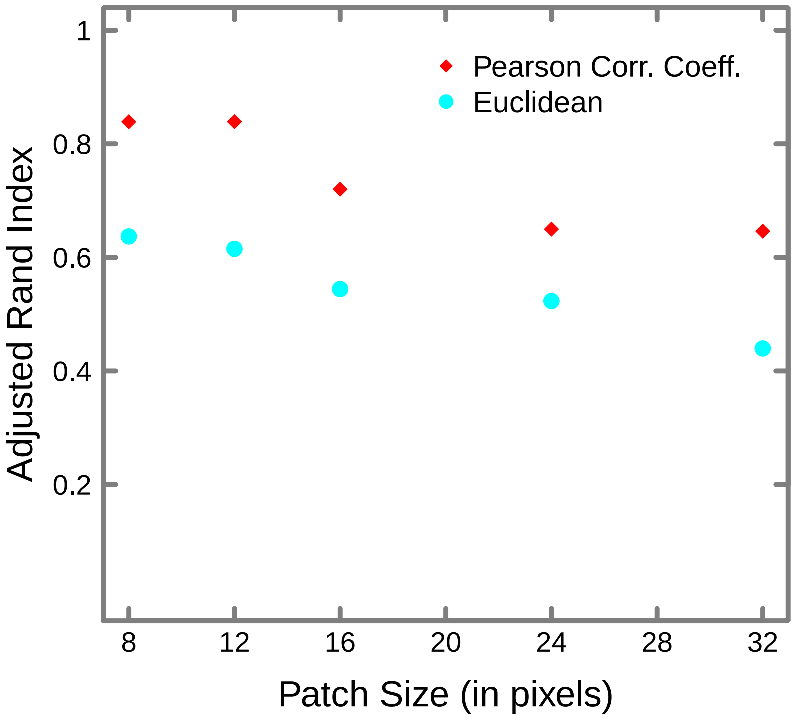


FIGURE 5.10: The performance of the similarity measure used when creating galaxy vector representations. This plot shows the effect on the clustering performance when using the pixel intensity power spectrum representation. The Pearson correlation similarity measure consistently showed higher performance across all representation types and applications.

5.6.5 Effect of Algorithm Used to Cluster Galaxies

I evaluated the effect of using two algorithms to cluster galaxies: agglomerative clustering using centroid linkage and K-means. K-means centroids were used to classify the galaxies in sky area 1. For agglomerative clustering centroid distance was used to classify galaxies. Figure 5.11 shows that agglomerative clustering generally outperforms K-means. It produces a more consistent result for different values of K. K is a pre-set parameter for the expected number of clusters. K-means performance is less robust for changes to K.

5.7 Summary

I present a method to cluster galaxies in telescope survey images. This method helps astronomers to identify specific types of galaxy without pre-classifying or pre-processing survey

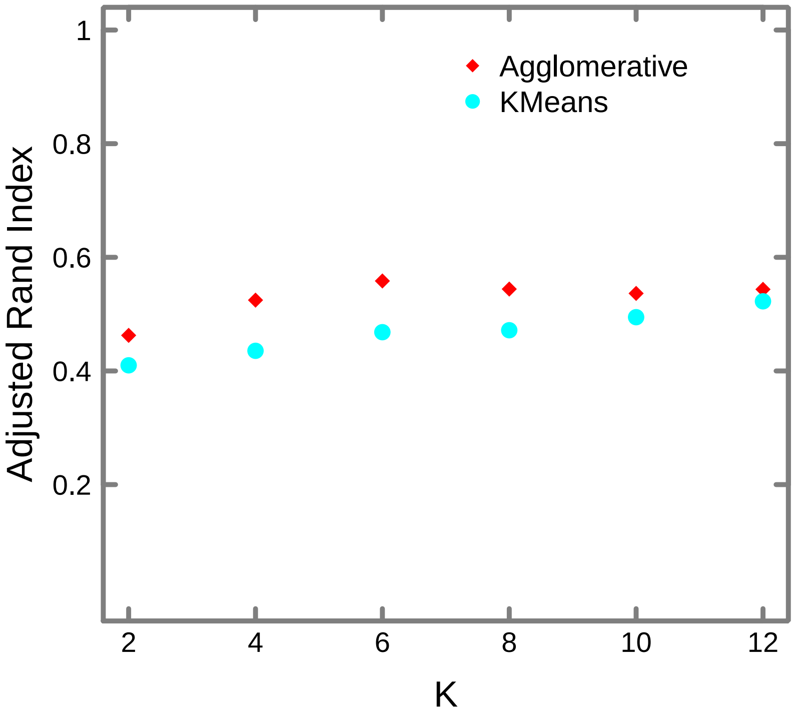


FIGURE 5.11: The difference between using agglomerative clustering and K-means for clustering galaxies using different values of K, a pre-set parameter for the expected number of clusters. The pixel intensity power spectrum representation was most effective for identifying late type and early type galaxies. Agglomerative clustering consistently out-performed K-means for all representation types and applications.

images. I demonstrate the effectiveness of the method by locating and clustering galaxies in large *HST* Frontier Fields survey images.

The method can detect different types of galaxies with the feature I call the pixel intensity power spectrum being the most effective at identifying elliptical, spiral and background galaxies. I find that the Rotationally Invariant Feature Transform (RIFT) is the most effective at detecting strongly lensed galaxies with an ARI of 0.48.

When testing parameters I show that patch size has a significant effect on two of the feature types with the pixel intensity power spectrum showing best results with small patch sizes, RIFT with larger patch sizes while Spin Intensity had stable results across patch sizes. When creating vector representations of galaxies I find the Pearson correlation measure to have a higher performance for all three features than when using the Euclidean distance. When clustering vector representations of galaxies agglomerative clustering out performed K-means.

In the next chapter I apply the method to the larger *HST* CANDELS survey Grogin et al. (2011) representing a data set size of 80Gb and many more types of galaxy.

Chapter 6

Application to Astronomy

The work in this chapter has been published in Hocking et al., 2018 *Monthly Notices of the Royal Astronomical Society*, 473, 1108.

6.1 Introduction

In the previous chapters I have developed a completely unsupervised technique. I have identified the most effective features, algorithms and hyper-parameters. The model selection process described in the previous chapter used typical machine learning performance measures for analysing the effectiveness of unsupervised algorithms. However, I have not yet used the technique to analyse any significant astronomy dataset nor evaluate the technique using standard astronomy measures. In this chapter I test the technique by cleanly separating late-type and early-type galaxies and analyse the results using colour-magnitude and M_{20} distribution diagrams. I then use the technique to analyse the five *HST* CANDELS fields. I analyse the galaxy groups using the 3D-*HST* photometry and photo-z data and I compare the groups to the Galaxy Zoo classifications for three of the CANDELS fields (Simmons et al., 2016b).

6.2 Identifying Late Type and Early Type Galaxies in the Frontier Fields

Galaxies classified as ellipticals (E0 to E7) and lenticulars (S0) are referred to as Early-Type galaxies, whereas those classified as spirals or irregulars are Late-Type galaxies. These terms were introduced due to an early interpretation of Hubble's Tuning fork that saw galaxies evolving from ellipticals on the left to spirals and irregulars on the right. This is no longer thought to be the case, but the terms have stuck.

The results in this section use images for strong lensing galaxy clusters Abell 2744 and MACS0416.1-2043 from the *HST* FF survey using the three filters: F435W, F606W and F814W. Chapter 2 describes this survey data.

6.2.1 The Learning Phase Applied to Frontier Field Abell 2744

6.2.1.1 Pre-processing

The input data matrix (DM1) consists of sample vectors that comprise a sequence of 8×8 pixel overlapping image patches sampled from each of the training images (the aligned F435W, F606W and F814W images of Abell 2744, Figure 2.2). Tests on various sizes of patches found that eight pixel square image patches produced the best results in terms of processing speed and galaxy detection; using larger patches resulted in a reduction in the identification of very small galaxies (effectively, this is a resolution issue). For each patch I evaluate the radially averaged power spectrum of pixel values in five bins, allowing us to encode information about the pixel intensity in a manner that is rotationally invariant. The power spectrum for each filter is concatenated into a single 15-element sample vector, that naturally encodes colour information to the data matrix. Thus the data matrix consists of rows of sample vectors and 15 columns called feature vectors.

To improve speed, during training I only consider regions of the image with pixel values in excess of 5 times the root mean squared value of blank sky in the image[1]. This reduced the number of image patches to 851,000. Note that these patches consist of small sections of galaxies and not whole galaxy images. Histograms of the feature vectors displayed log normal distributions.

[1] Although note that in principle this data could be used during training

In order to convert each feature to a normal distribution, thus creating a better clustering outcome, I simply took the natural log of values in the data matrix. Each of the feature vectors were then normalised by subtracting the mean and dividing by the unit of standard deviation.

6.2.1.2 GNG & Hierarchical Clustering

I configured the maximum nodes parameter of the GNG algorithm (Section 3.3.2) to 20,000 and processed each of the 851,000 sample vectors 100 times. The output of this step is data matrix (DM2) of $20,000 \times 15$, representing the code vectors of the GNG nodes. DM2 is then used as input into the HC algorithm (Section 3.3.3). The HC was run with three types of similarity measure including: Euclidean distance metric, cosine similarity measure and the Pearson correlation coefficient, with the Pearson correlation coefficient (see equation 3.2) achieving the best results. I searched down the resulting hierarchical tree from the root node to identify the relevant child groupings (clusters) of GNG nodes. Each cluster contained a corresponding error value which indicated the 'quality' of the cluster. I selected all the clusters that had an error of 0.15 or less, which identified 536 independent clusters of GNG nodes. Using a higher error value would identify fewer clusters that contained larger numbers of GNG nodes. However, the next steps in the process are not sensitive to larger numbers of clusters and therefore I chose a smaller error value which represented higher quality clusters that are more accurate (i.e. the GNG nodes are more similar). Using GNG and HC I have identified 536 groups that contain the original population of 851,000 patches.

6.2.1.3 Connected-component labelling

I used the connected-component labelling algorithm described in (Section 3.4.2.1) to identify spatially connected sub images (components) in DM1. These connected overlapping patches represent the individual galaxies. The Frontier Fields images contain crowded central fields with bright, extended stellar halos around elliptical galaxies. In order to separate the galaxies in the central elliptical cluster I identified two thresholded lists of the 851,000 overlapping patches. One list identified the sub images at locations with pixel intensity of at least 5σ over the background level and a second list identified the overlapping image patches at least 10σ over the background level (where 1σ is the root mean square value of the source-free background). The locations of the 851,000 overlapping patches were identified and the mean pixel intensities from each of the three bands were compared to the threshold level. If any of the pixels were over the threshold level in any of the three bands the image patch was added to the list.

The connected-component labelling process used the following inputs i) the co-ordinates of each of the 851,000 image patches ii) the size of the image patches (8×8 pixels) iii) a minimum component size of five, so that only components with five or more overlapping patches were considered, and iv) the 5σ and 10σ threshold lists. Any component overlaps were identified and the 10σ component was selected in preference to any overlapping 5σ component. This enabled the galaxies in the brightest areas of the extended stellar halo to be distinguished. A catalogue of the components was created by calculating the approximate position of the component (calculated using the average position of its sub images) and the width and height of the component was calculated by identifying the minimum and maximum co-ordinates of the sub images.

6.2.1.4 Identifying galaxies

The next step combined the 536 clusters of sub images from the HC process and the components identified by the connected component labelling process to create a new data matrix (DM3). Each sample vector in the data matrix represented a component (galaxy) consisting of 536 elements. The value of each element was a count of the number of overlapping patches in the component that was in the representative hierarchical cluster. The resulting sample vectors were sparse in that the majority of the elements were zero. The final preparation step used to create DM3 was a normalisation: divide each element in the sample vector by the sum of all its elements. A large galaxy and small galaxy of the same type will consist of the same types of patches (identified by HC). However, there will be a large difference in patch counts in each element. Therefore I divided each element in a sample vector by the sum of the vectors elements which rescaled the vector elements to de-emphasise galaxy size.

The final step was to use HC again on DM3 to identify ‘clusters’ of galaxies that are similar to each other, using the cosine similarity measure. Cosine similarity is a measure of the angle and not magnitude between two vectors and therefore improves the scale invariance of the process. I ran the algorithm with a pre-set parameter (K=2) to output two clusters or groups of galaxies.

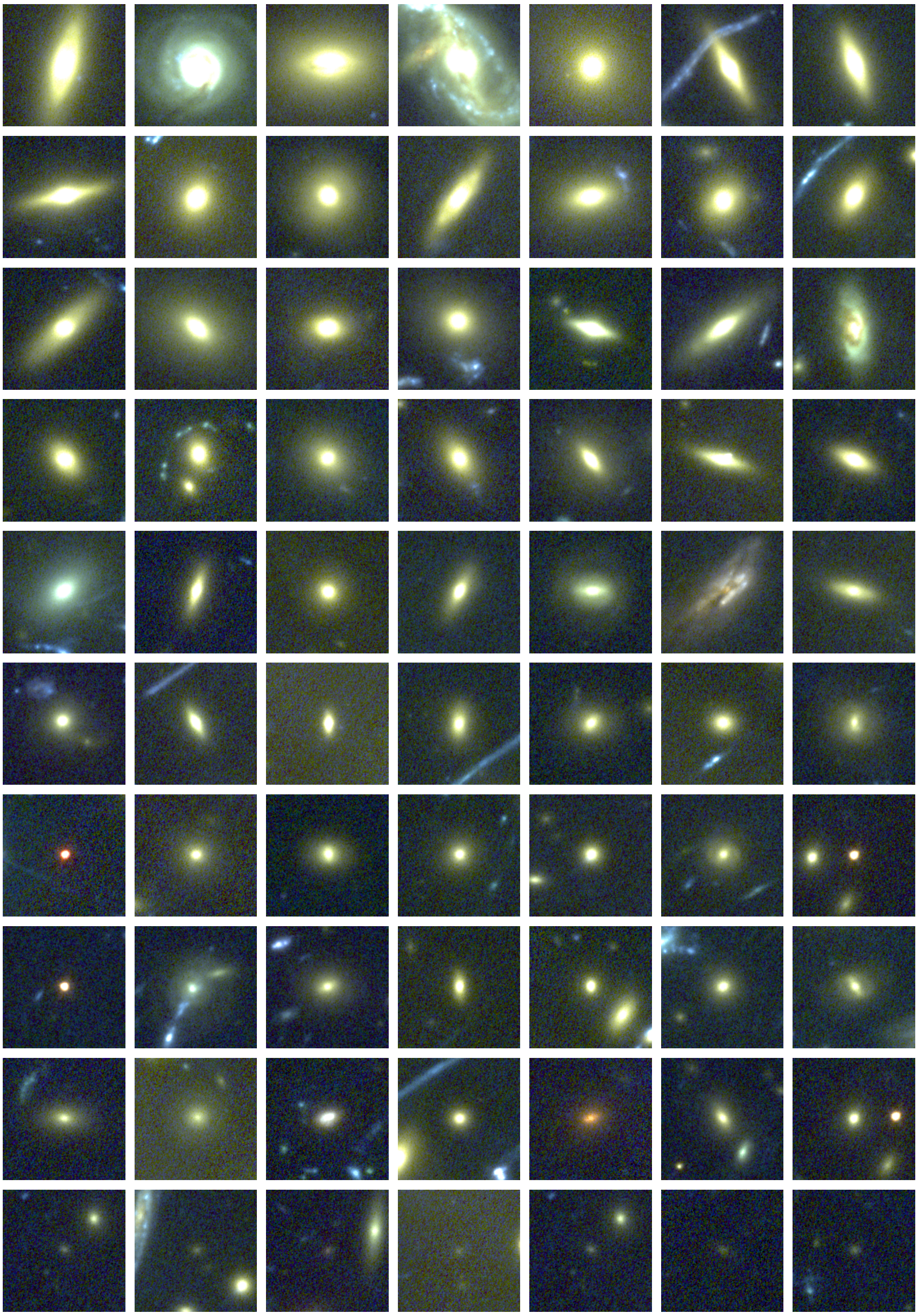


FIGURE 6.1: Examples of a sample of galaxies in MACS0416.1−2403 that the algorithm automatically identifies as being members of group ‘one’. Each image is $4.5'' \times 4.5''$. The algorithm automatically identified this group and classified these galaxies using no data other than the image pixel intensity values from the F435W, F606W and F814W bands, and based classifications on the information in the Abell 2744 image.

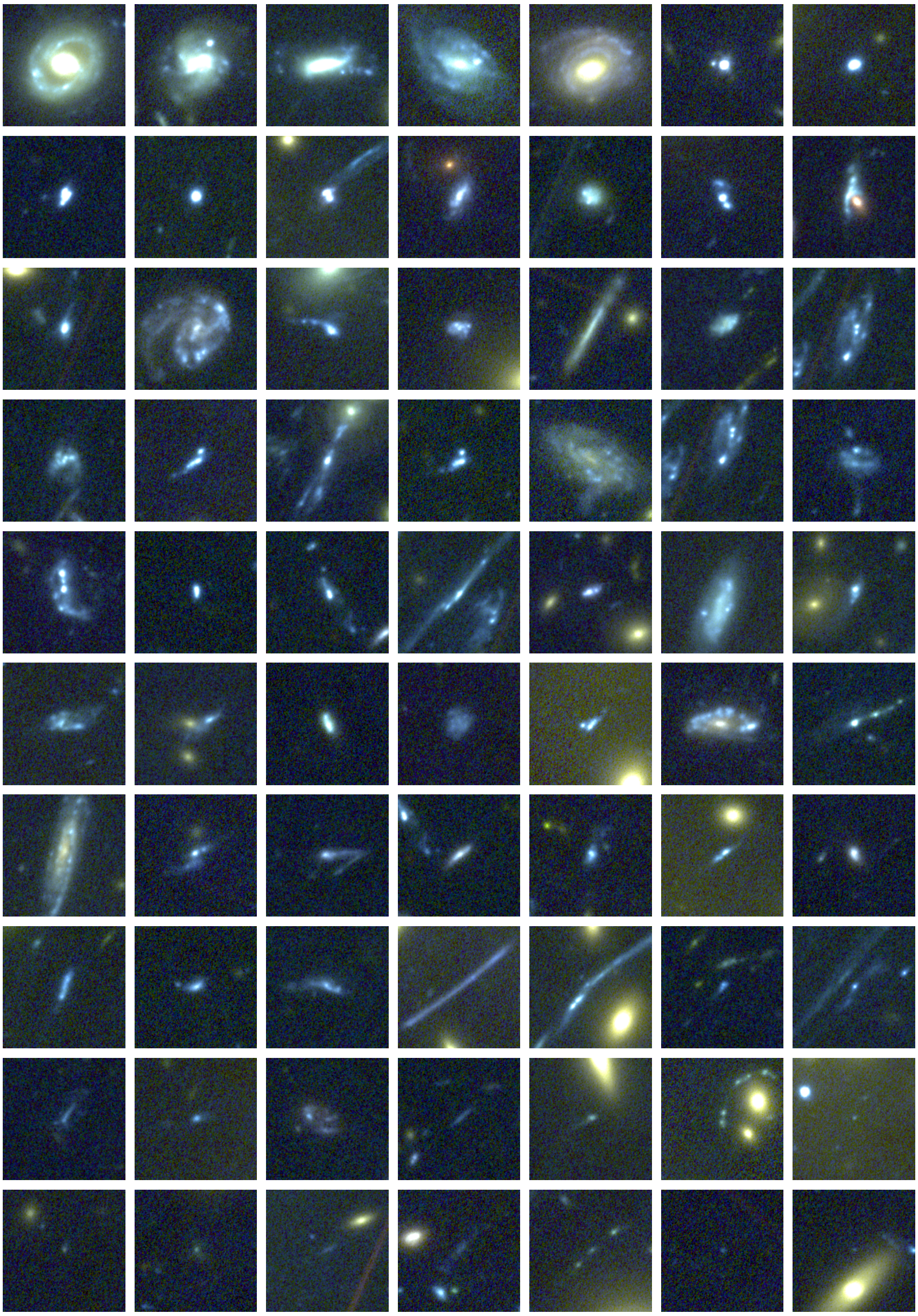


FIGURE 6.2: Examples of a sample of galaxies in MACS0416.1−2403 that the algorithm identifies as being members of group 'two'. Each image measures $4.5'' \times 4.5''$ arcseconds. Lensed galaxies are included in this group. Again, the algorithm automatically identified this group and classified these galaxies using no data other than the image pixel intensity values from the F435W, F606W and F814W bands, and based classifications on the information in the Abell 2744 image. Note that in some cases the algorithm has correctly classified faint galaxies that are clearly in the stellar halo of an elliptical.

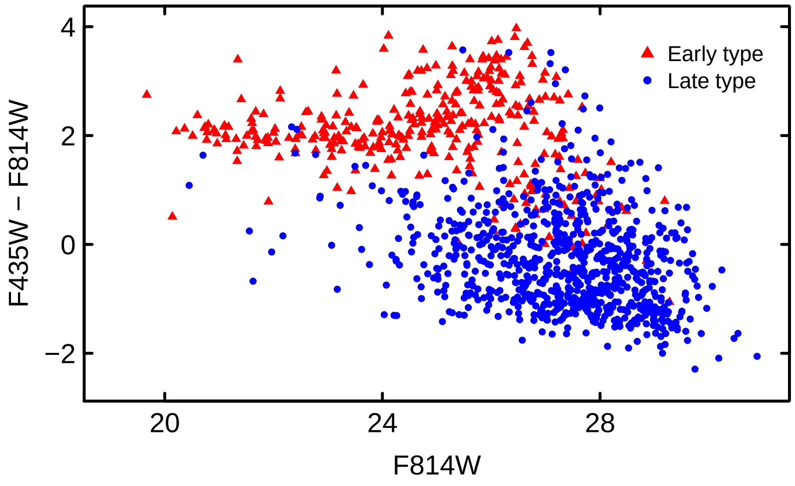


FIGURE 6.3: A colour-magnitude diagram of the galaxies in MACS0416.1−2403. The galaxies that the process identifies as being members of group 'one' are labelled with the red triangles. The galaxies that the process identifies as members of group 'two' are labelled with blue circles. The process cleanly separates the early types in the red sequence and the late types in the blue cloud.

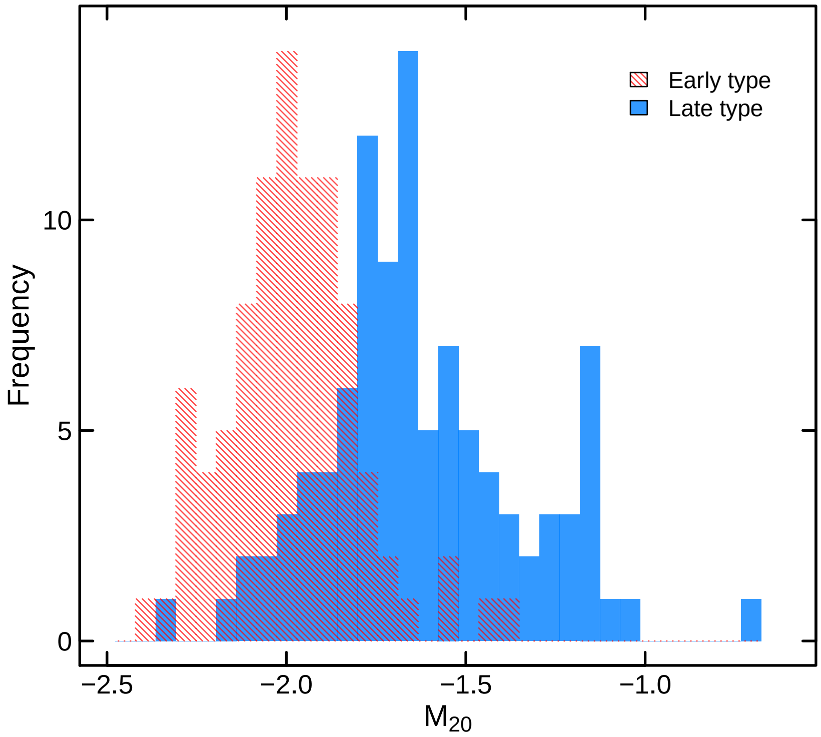


FIGURE 6.4: Histograms showing the M_{20} morphological measure calculated for the galaxies that the process identifies as being members of group one in red, and the galaxies that the process identifies as being in group two in blue. This appears to identify two populations of galaxies as found in Lotz et al. (2004).

6.2.2 Verifying the method on Frontier Field MACS0416

The learning phase identifies two groups of galaxies in the Abell 2744 images broadly representing late type (blue spiral, irregular, lensed) and early type (red, smooth, elliptical) galaxies. I then used the trained network to analyse a new, unseen image of the same type (MACS 0416.1−2403). This analysis identified the same two groups of galaxies and produced a catalogue of the galaxies and their type. Example galaxies from these two groups are shown in Figures 6.1 and 6.2. No pre-existing labels are available, therefore typical measures used in supervised machine learning to analyse accuracy such as precision/recall are not available. Instead, in order to verify the results, I investigate how the method compares to two traditional techniques for classifying early/late type galaxies. First, the two classes of galaxy should be cleanly separated in a colour-magnitude diagram (as explained in Section 2.3), and indeed I find this is the case (Fig 6.3.). Aperture photometry was measured using SExtractor on cut-outs of each galaxy in the classified sample. The figure shows the algorithm correctly identifies the red sequence and blue cloud, although clearly with some scatter between the point clouds; generally these are due to close blends and projections. I also calculated the M_{20} morphological parameter Lotz et al. (2004) for galaxies larger than 15×15 pixels in the F814W band. M_{20} is the normalized second order moment of the brightest 20% of the source flux, with less negative values corresponding to clumpier sources. Figure 6.4 shows the results, which shows a systematically lower M_{20} value for our early types compared to our late types. I argue that Figures 6.1–6.4 demonstrate the proof-of-concept success of the algorithm in automatically classifying sources into astrophysically meaningful groups. In the following I apply this method to a broader input set – the *HST* CANDELS fields.

6.3 An Automatic Taxonomy of CANDELS Galaxy Morphology

In this section I describe how the technique was used to categorise galaxies in the *HST* CANDELS dataset which is described in Chapter 2.

The process to analyse the FF images in Chapter 5 used generalisation by training on one field and then applying the model to classify objects in a second field. I took this approach as it was important to prove that it is possible to do this using an unsupervised approach in order to significantly reduce processing time as the computational time for applying GNG and HC on very large data would be prohibitive. However, when considering the size of the CANDELS dataset

(F160W and F814W imaging) it is only ∼60 Gb and therefore I apply the learning algorithms to all five fields of the CANDELS data in its entirety.

Before describing the CANDELS classification process I point out that combining the data from the deep and wide fields is not ideal for machine learning processes. The initial assumption is that data is prepared in a consistent manner. In this case, the depth of the images varies across the fields and in some cases the classification process identifies groups that contain galaxies predominantly from GOODS-North and GOODS-South and other groups predominantly from UDS, EGS and COSMOS. In §6.3.2 I compare our catalogue to the Galaxy Zoo: CANDELS classifications and I note that the Galaxy Zoo team has provided alternative weighted classifications for the galaxies in the deep sections of the survey, illustrating that the combination of depths appears to affect human classifiers too.

The first step is to select the pre-set parameters. It was unclear whether the pixel scale and reduced depth compared to the FF images would affect the parameter choices therefore I ran the process multiple times with different options such as two patch sizes (8 and 12 pixels) and two threshold levels (4σ and 5σ). On inspection of the results I chose a patch size of 12 pixels and a threshold level of 4σ above the background level. This produces 9.5 million overlapping patches, each of which were then normalised and topologically mapped by GNG to 10,000 GNG nodes. I applied the HC algorithm using the Pearson correlation which resulted in 1,174 groups (using a threshold of 0.045). I select the threshold level based on the quantisation error of the patch groups.

For each of the five fields the connected components step was run to identify galaxies and create the galaxy vector representations that are then grouped together by another HC step using the Pearson correlation. The output is a hierarchy of galaxy classifications. At the top level I choose a minimum number of clusters of 100 and then for each level increment by twenty until the lowest level contains 200 distinct classifications. In addition, I calculate an 'average' galaxy vector representation in each group by averaging the vector representations of all the galaxies in a group. A similarity value between each galaxy and the 'average' galaxy for its group is calculated by computing the Pearson correlation between the vector representations and subtracting it from one. The most similar galaxies will have similarity value of 0. Note that any negative correlations are heavily penalised. This value is important to identify the purest examples in each classification. I provide the similarity scores in the catalogue as 'classification distance' and it is important to use these values to sort the galaxies in each classification.

TABLE 6.1: The format and columns of the catalogue produced by the machine learning technique.

Column Position	Column Name	Description
1	Field Id	The identifier of the field where the object resides. 0 GOODS-N, 1 UDS, 2 EGS, 3 COSMOS, 4 GOODS-S
2	Object Id	The ID of the object from the 3D *HST* catalogue by Skelton (based on a cross match)
3	RA (degrees)	Right Ascension (J2000)
4	Dec (degrees)	Declinaton (J2000)
5-10	Classifications	Hierarchical classifications, 6 levels of classifications
11-16	Classification distances	A number between 0 and 1. The nearer to 0 the more relevant the galaxy is to the classification. These fields are important for sorting objects within classifications.

Choosing the number of clusters is one of the main difficulties of the technique. I have selected a range from 100 to 200 clusters using visual inspection of the classifications to identify which levels create the purest classifications. On inspection of the results the higher granularity of 200 clusters appear to provide the purest classifications. Fortunately the use of the hierarchical clustering algorithm makes it straightforward to retrieve different numbers of clusters without requiring significant re-processing. The larger number of clusters required for CANDELS may be due to this being a higher redshift survey with a significant variation in galaxy morphology. Or, it could be a bias of the algorithm that it requires a larger number of clusters to work effectively. For other datasets consisting of large numbers of different morphologies it is likely that using a similar value of 100 to 200 clusters is a good starting point. However, a dataset with less variation in morphology may require a smaller number of clusters.

The catalogue provides classifications for $\sim$60,000 galaxies. Table 6.1 contains the description of the classification catalogue file. The catalogue file is available in CSV format and I provide a visual version of the catalogue at www.galaxyml.uk. In addition, as each galaxy has a vector representation I can also use the Pearson similarity measure to identify the most similar other galaxies within CANDELS for each galaxy. I have used this capability to provide a web based galaxy similarity search function at www.galaxyml.uk.

To use the catalogue it is important to employ the classification distance column to sort all the galaxies in ascending order. The classification distance columns are shown in Table 6.1.

These distances identify how close each galaxy is to its particular classification. The higher the classification distance the less similar a galaxy is to the classification it is member of.

In order to analyse the classifications and to produce the final catalogue I matched my classification catalogue to the 3D-*HST* catalogues from Skelton et al. (2014), which contain photometry and photometric redshifts for CANDELS. Skelton et al. (2014) determined the photometric redshifts by using EAZY (Brammer et al., 2008).

Figure 6.5 shows colour-colour plots for some example galaxy classifications, and illustrates the effect of hierarchy: each top level group is split into further levels, which are sub-sets of higher levels. They can be considered increasing levels of detail. Different classifications tend occupy distinct regions of colour space, and it is clear that the stellar locus (a region on a colour-colour diagram where stars are located, often used to separate stars and galaxies) is clearly delineated. This is not surprising, since colour information is encoded in the classification process (albeit a single colour in this case). Figure 6.6 shows the F606W total magnitude distributions for selected classifications where again I can see that automatically classified groups tend to have well defined magnitude distributions distinct from the overall population. Finally, photometric redshift distributions are shown in Figure 6.7; again, showing well defined distributions for different automatic classifications. This demonstrates that the algorithm is actually grouping sources together that can be linked to (or labelled with) well-defined and well-understood observed parameters, and therefore can be put into a practical astrophysical context. Figures 6.8, 6.9, 6.10, 6.11 are examples of galaxies within different groups and levels within the hierarchical catalogue, illustrating how the algorithm is grouping together similar types of object over a wide dynamic range. While the majority of the classification groups appear well-defined, I note, however, that not all the classification groups are clean. Three examples are shown in Figure 6.12. They contain inconsistent galaxies, galaxies near the edge of coverage and also galaxies that appear to be outliers.

The redshift groupings in Figure 6.7 may reflect physical differences in galaxy populations at different redshifts due to evolution, or simply the effect of observing similar galaxies at different distances, i.e. the impact of redshift on observed colour and size. It may be possible to differentiate between these two options through SED fitting. This is a technique that requires observations at multiple wavelengths, which are then compared to different models to obtain estimates of physical properties, such as mass and star formation rate. If galaxies in different classification groups were found to have different mean physical properties, then this could

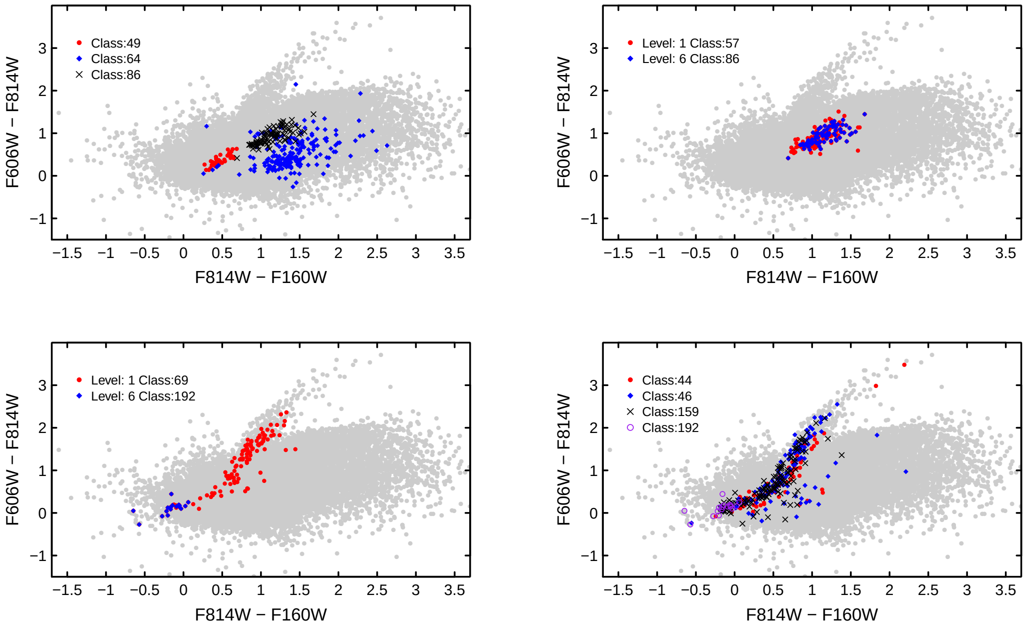


FIGURE 6.5: These colour-colour diagrams show some of the classification groups in our classified CANDELS catalogue. The background grey points are a random sample of the entire population. In blue, red and black are galaxies from individual classifications. Many of the classifications appear as distinct clusters in colour-colour space. The top right shows galaxies from classification number 57 and one of its 'child' classifications number 86 which is an example of the hierarchy within the catalogue. The bottom left figure also shows the effect of the hierarchy of classifications, level six being the most detailed classification level, and level one at the highest (coarsest) level. The bottom panels show different classifications for point sources which track the stellar locus; note that in the bottom right panel I find different classifications for sources lying in the same colour space, indicating that, while colour information clearly enters into the classification, the algorithm can offer a more finely controlled object classification and selection.

suggest that an evolutionary difference has been detected rather than simply an observational effect.

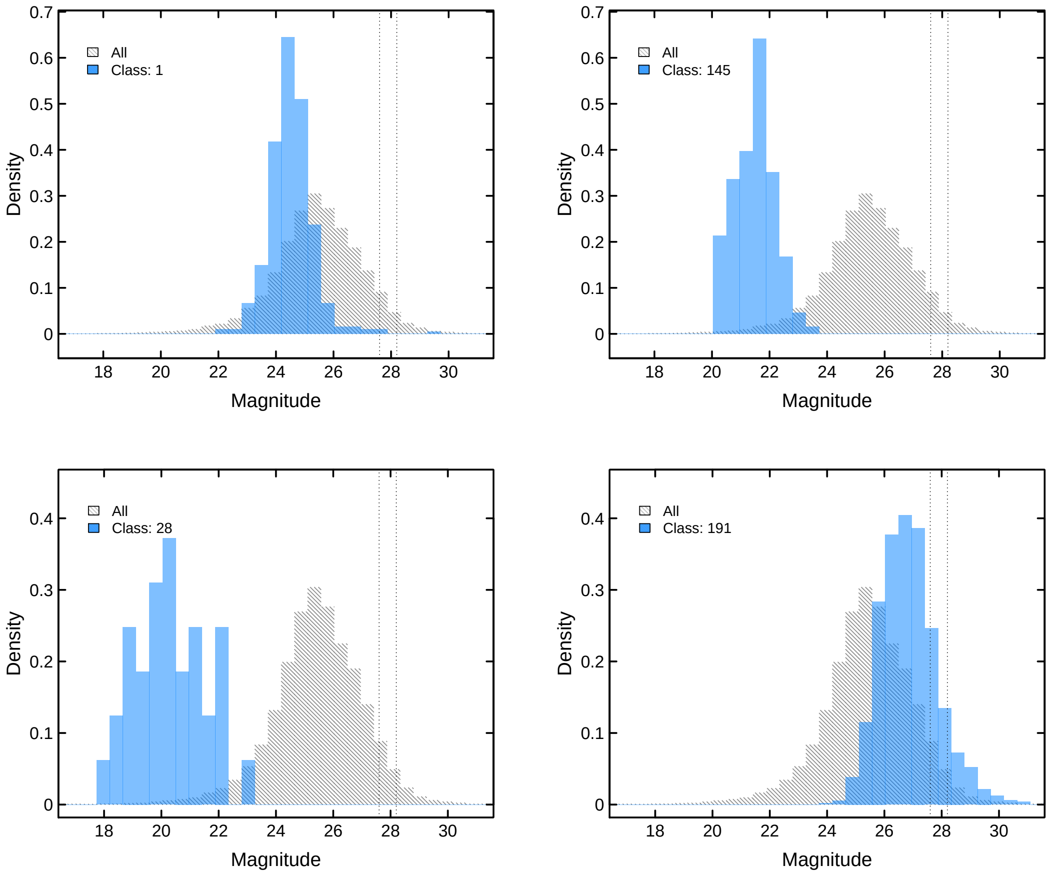


FIGURE 6.6: These histograms show F606W total magnitudes obtained from the 3D-HST photometric catalogues of Skelton et al. (2014). They compare the magnitude distributions of galaxies given a specific classification (blue) with a random sample of galaxies from the full entire population (grey). The vertical lines are the 5σ limiting magnitudes for the wide and deep CANDELS surveys. This figure illustrates that the classification process groups galaxies into categories that can be easily described in terms of traditional descriptors such as magnitude, with distinct and 'well behaved' distributions.

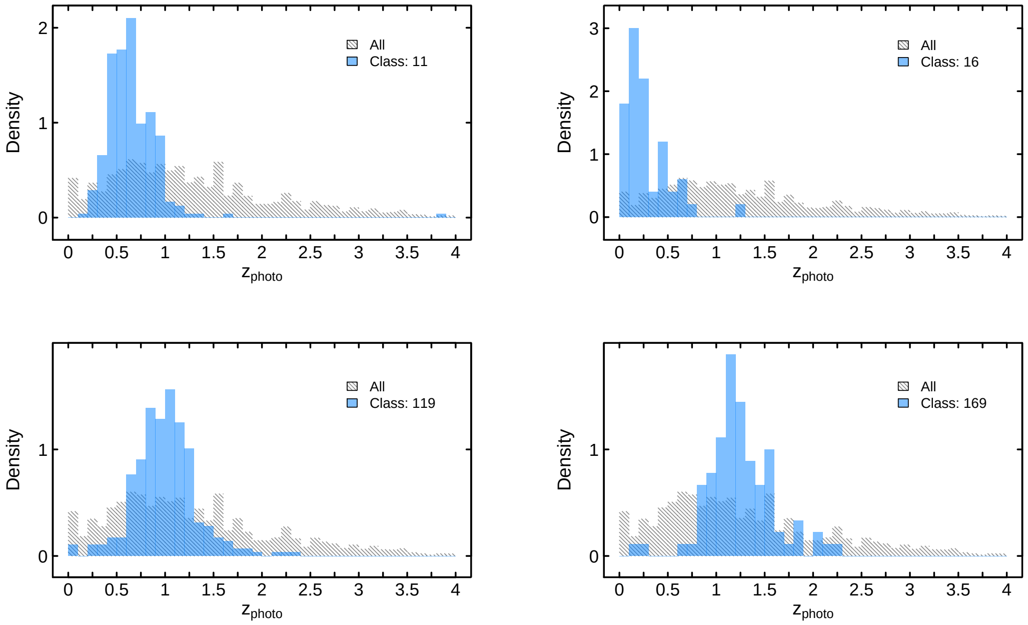


FIGURE 6.7: I show the photometric redshifts of the galaxies for four different classifications identified by the machine learning technique. The photometric redshifts were obtained from Skelton et al. (2014) who determined them by using EAZY (Brammer et al., 2008). The histogram in grey shows the distribution for a random sample of the full population. As in Figure 6.6, each classification (based solely on pixel data) falls into well behaved distributions; for example, class 119 (bottom left) clearly contains galaxies at $z \approx 1$. Adding these 'post-processed' labels to automatically classified sources is useful in assigning astrophysical context to the groups the algorithm has identified.

The catalogue was used to create Figures 6.8, 6.9, 6.11 in the following way. For each field the FITS files for F160W, F814W and F606W were combined into a single PNG image file using STIFF (Bertin, 2012). The catalogue file was then used to identify galaxies in each classification. The galaxies were sorted by their classification distance in ascending order. For example, Figure 6.11 includes three rows from classification 169. The field, RA and Dec for each galaxy were extracted from the catalogue file. The galaxies were sorted using the classification distance in ascending order. The pixel co-ordinates were identified using the FITS file header and an image patch was cut from the field PNG file around the galaxy. The visual version of the catalogue on the website www.galaxyml.uk showing all 200 classifications for the most detailed classification level was also created using this method.

Figure 6.10 was produced by manually selecting and ordering ten classifications (23, 174, 6, 86, 45, 8, 11, 140, 30, 146) from the website visual catalogue. These classifications were selected to demonstrate the granularity of classification that is possible using the technique. The catalogue was then used to create the images by repeating the process used for Figures 6.8, 6.9, 6.11.

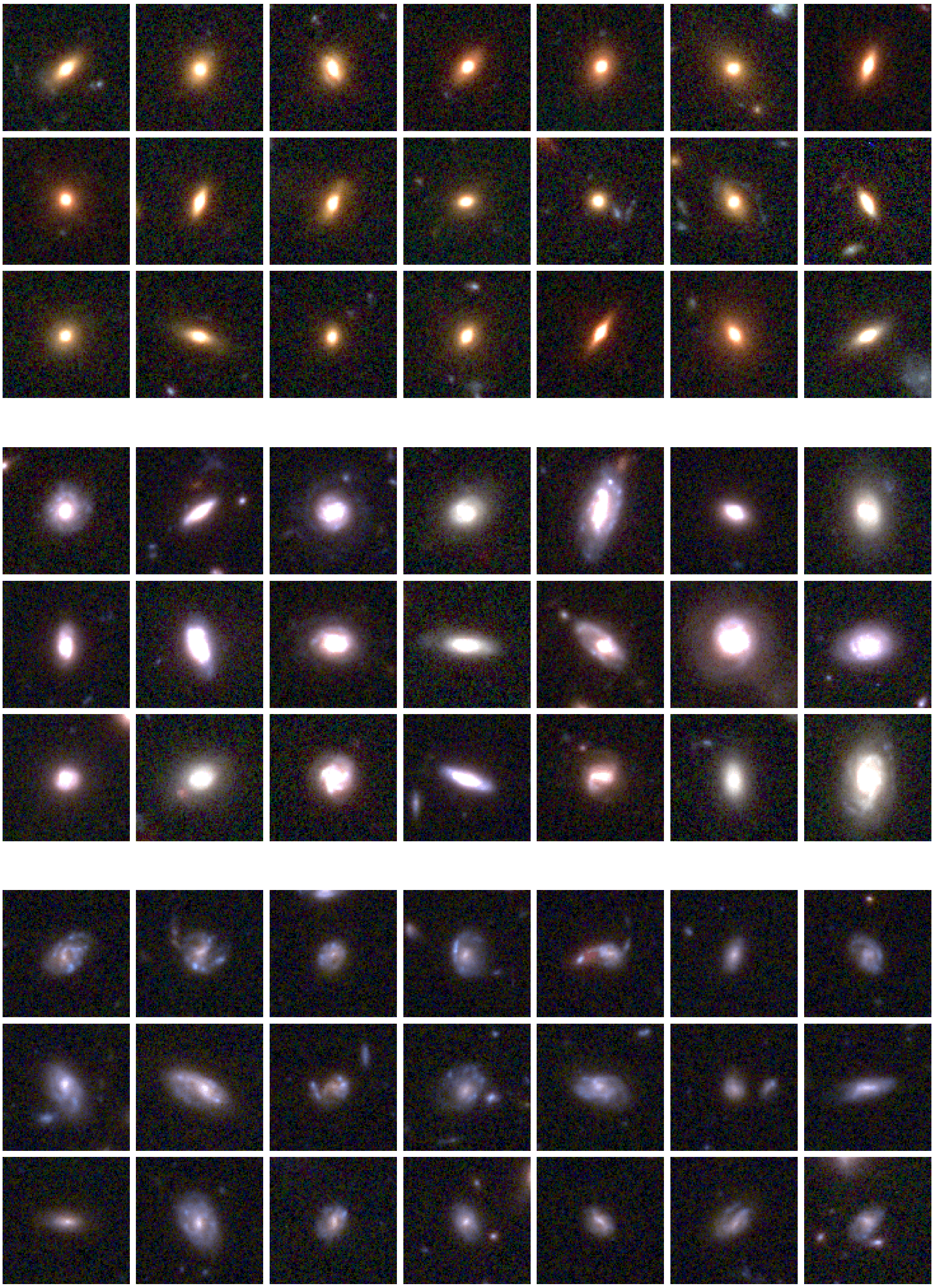


FIGURE 6.8: Example images from the top level of three different CANDELS classification groups (classification groups 7, 18 and 98). Each image is $6'' \times 6''$. The galaxies in each group are ordered row-wise in order of their similarity to the 'average' classification in the parameter space of the group. The top left image is the most similar galaxy to the 'average' and the bottom right is most dissimilar. The classification catalogue provides these as classification distances which can proxy as a quality flag. The distances are normalised between 0 and 1, with 0 being an identical match to the average. Here the RGB channels are the F160W, F814W and F606W bands, but note that the latter was not included in the learning.

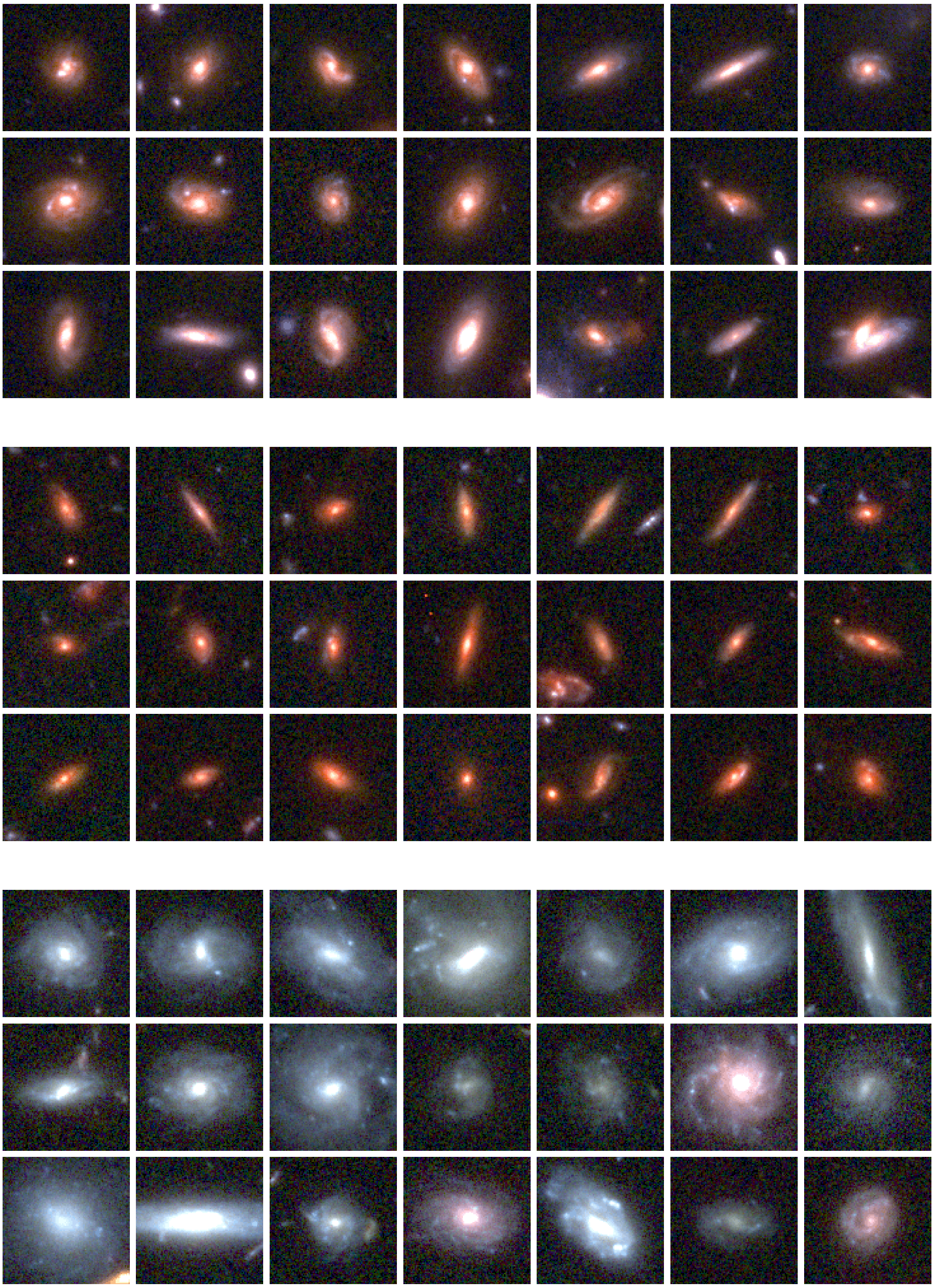


FIGURE 6.9: Examples of galaxies in three classification groups (30, 36, 48) from level one (the coarsest classification) in the hierarchy. As before, each image is $6'' \times 6''$ and ordered left to right in order of similarity to the 'average' galaxy in the group. The RGB channels are the F160W, F814W and F606W bands.

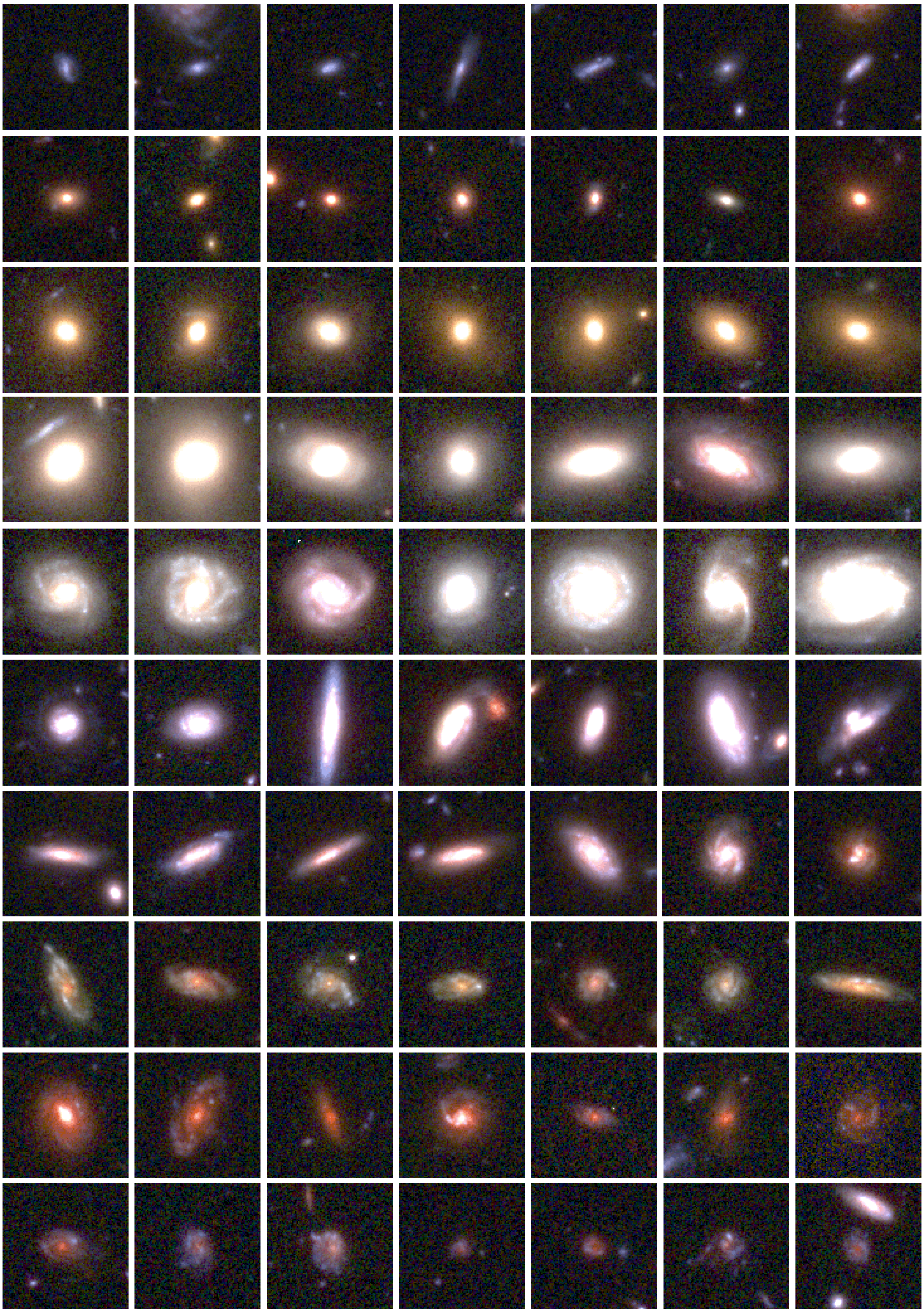


FIGURE 6.10: Each row of $6'' \times 6''$ images shows galaxies in an individual classification group, and are selected from the lowest hierarchy level in that group. The galaxies are ordered left to right by their similarity to the average galaxy with the first panel most similar to the average. Again, the similarity of sources in each group is clear. The RGB channels are the F160W, F814W and F606W bands.

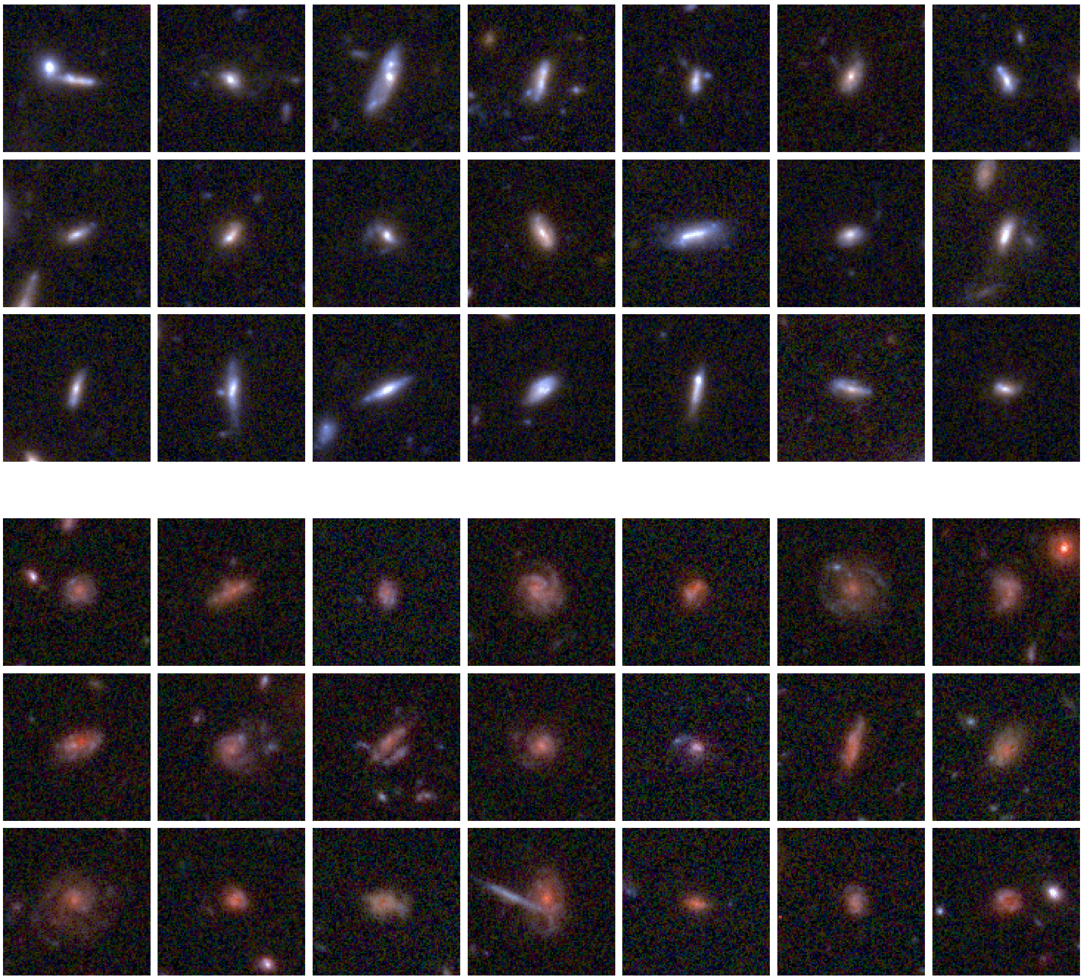


FIGURE 6.11: Examples of galaxies in two classification groups: group 8 at level one (low level of refinement) and group 169 at level six (higher level of refinement). The images are $6'' \times 6''$and the RGB channels are the F160W, F814W and F606W bands.

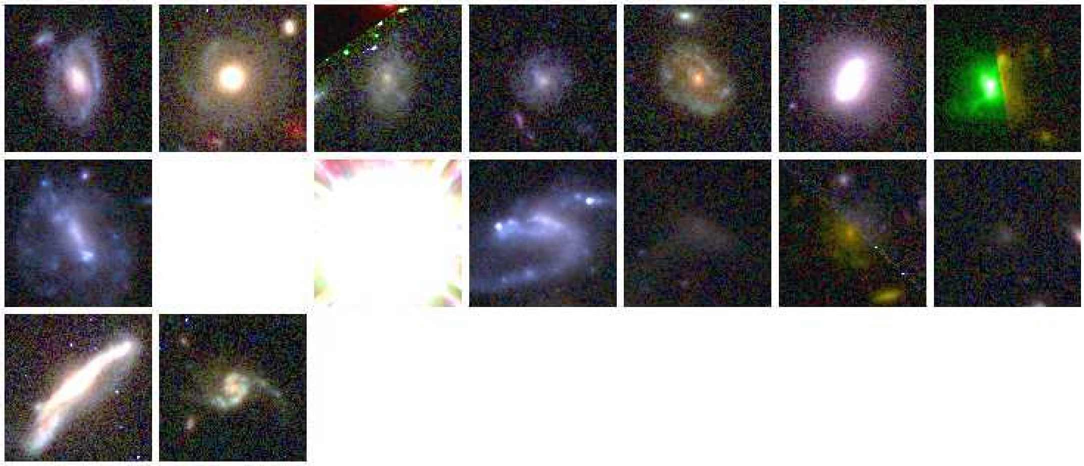


FIGURE 6.12: The majority of the classifications groups are very clean. However, there are some that are less so such as these three classification: 24, 41 and 56. Each row is an individual classification. The third row appears to include objects that are outliers distinct from other galaxy classifications.

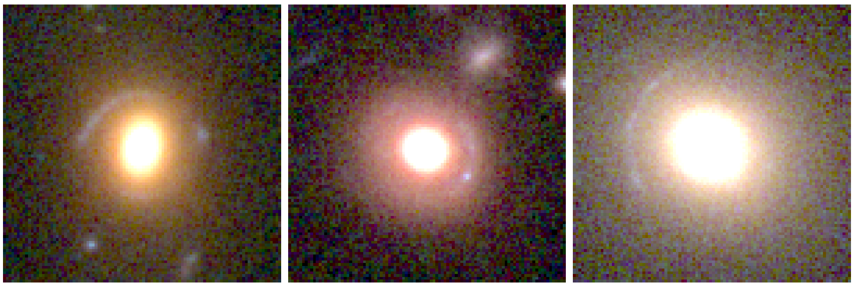


FIGURE 6.13: Two potential strong lensing candidates (left and middle) and a known lens (right). All three appear in the same classification group. The galaxy to the left is in UDS, ID16074 at location $02^{h}17^{m}06\overset{s}{.}2$(RA) and $-05^{\circ}13'17\overset{''}{.}6$(Dec.). The galaxy in the middle is in EGS, ID10397 at $14^{h}19^{m}00\overset{s}{.}12$(RA) and $+52^{\circ}42'48\overset{''}{.}9$(Dec.). The known strong lensing galaxy COSMOS 0013+2249 is shown on the right.

6.3.1 Identifying Unusual Objects

This technique can be used to identify rarer types of object. An advantage of the technique is that we can use different algorithms in place of hierarchical clustering to achieve a different view of the survey images. One such algorithm is K-Means (Sculley, 2010) which can be used in place of hierarchical clustering. I have explored variations of parameters and algorithms in Hocking et al. (2017). I analysed CANDELS with a variant of the machine learning system using K-Means. I scanned the classification groups to identify which groups contained galaxies with large elliptical central bulges but with localised higher emission in the shorter wavelengths - this could be the result of mergers or conjunctions with background galaxies. I identified a strong lensing galaxy that is currently known in the NASA Extragalactic Database (NED)[2], COSMOS 0013+2249 at a spectroscopic redshift of $z = 0.3461 \pm 0.001$ (Faure et al., 2011). In addition, I found two candidates which are not classed as lenses in NED. These are ID16074 at $z = 0.628 \pm 0.007$ and ID10397 at $z = 0.389 \pm 0.007$, IDs and photometric redshifts are from Skelton et al. (2014). All three lenses are shown in Fig 6.13.

6.3.2 Comparison to the Galaxy Zoo CANDELS Classifications

Galaxy Zoo (GZ) has been providing crowd-sourced statistically robust visual morphological classifications for some years now (Lintott et al., 2008, 2011; Willett et al., 2013). They have turned their attention to CANDELS and have recently published detailed morphological classifications for three of the CANDELS fields: GOODS-South, UDS and COSMOS. The classifications were provided by 95,000 volunteers with each galaxy receiving an average of 43

[2] The NASA/IPAC Extragalactic Database (NED) is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

classifications (Simmons et al., 2016b). GZ leads a volunteer agent through a decision tree of questions about an individual galaxy. Depending on the answer to a question the agent will follow different paths down the decision tree. Between two and nine questions are asked of the volunteer; for example, if the galaxy image is a star or artifact then only two questions are required, if it is a spiral galaxy then up to nine questions are required. These classifications have been consolidated and robustly analysed by the GZ team to provide a catalogue of weighted fractions for each answer in the tree for 48,000 galaxies. Simmons et al. (2016b) describe the catalogue, the methodology and provide a detailed analysis.

How do the machine learnt classifications and the human classifications compare? Clearly I cannot expect a direct mapping between GZ classifications and my hierarchical grouping, but I can use the GZ catalogue to ask the question of whether our groupings would have had a 'concordance' classification (based on the questioning tree) from a cohort of human classifiers. GZ provides two catalogue files. One is the full catalogue for COSMOS, UDS and GOODS-South and a second that contains adjustments to the classifications made to the deep survey for GOODS-South. I choose to compare our data with the original catalogue as no adjustment has been made to the machine learning technique for the different depths. The catalogue files provided by GZ include the number of classifications and the weighted and unweighted fractions for each answer in the decision tree. I consider three top-level questions: T00 A0 '*is the target smooth and rounded?*', T00 A1 '*does it contain features or a disk?*' and T00 A2 '*is it a star or artifact?*'. The weighted fractions run from 0 to 1 corresponding to a negative or affirmative result. I ask whether the algorithm has assembled groups for which $\geq$50% of the members (that have GZ classifications) have answers to any of these questions above a weighted fraction of 0.5. We call this a 'concordance' classification. I can find several examples of concordance classifications, and I show two examples of each in Figures 6.14, 6.15 and 6.16, presenting the top seven galaxies from each classification as a guide, and the distributions of the weighted fractions of the answers to each of T00 A0–2 for galaxies in each group.

GZ also includes several 'clean' classifications, where a boolean flag is assigned to a subset of GZ classifications for which the weighted classification indicated a high conviction for 'clean_feature' (229), 'clean_spiral' (278), 'clean_smooth' (4662), 'clean_edge_on' (162) and 'clean_clumpy' (215). The numbers in parentheses indicate the number of clean classifications in the matched catalogue. Note that the majority are 'clean_smooth'. The top level of our hierarchical classification contains 100 groups, and can be considered the coarsest level of classification refinement. This is probably most suitable for this comparison: I can simply assess

TABLE 6.2: Fraction of top level machine learnt (ML) classification groups containing 50% and 100% of the galaxies in the various Galaxy Zoo 'clean' classes.

Galaxy Zoo clean class	Fraction of ML groups containing 50% of clean class	Fraction of ML groups containing 100% of clean class
Smooth	13%	89%
Spiral	7%	52%
Featured	5%	53%
Clumpy	8%	51%
Edge-on	5%	47%

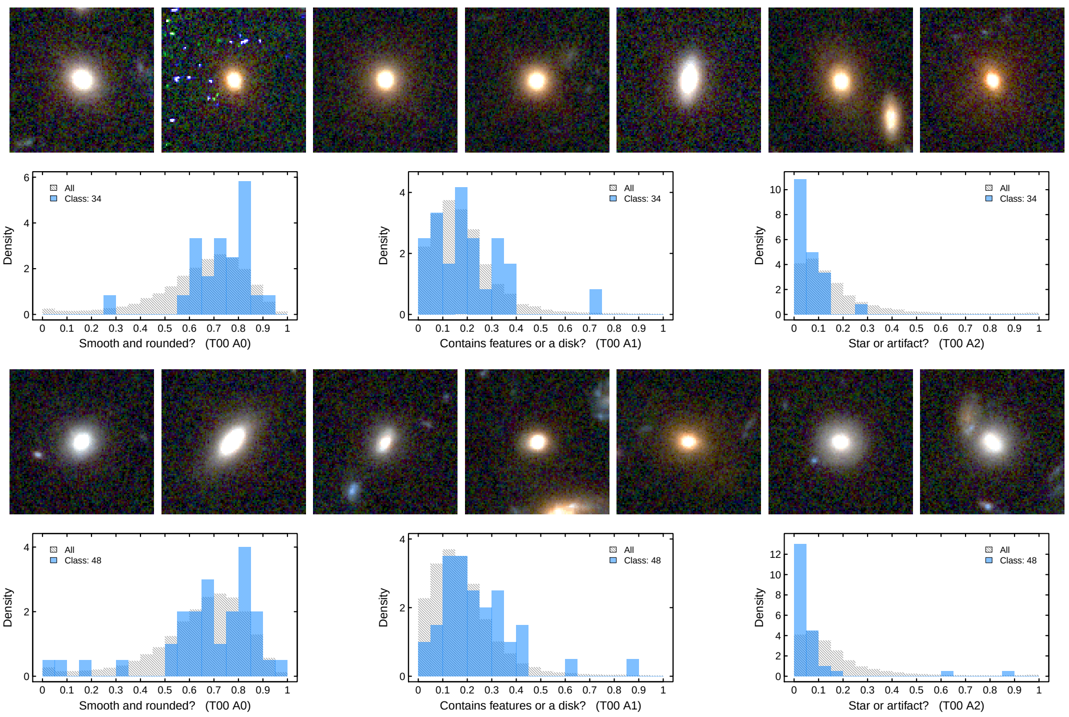


FIGURE 6.14: This figure shows two examples of what I refer to as a 'concordance' group, where over 50% of the galaxies for which Galaxy Zoo classifications were made have a weighted fraction over 0.5 for question T00 A0 '*is the target smooth and rounded?*'. The images show the top seven matches in the group and the histograms compare the distribution of weighted fractions for questions T00 A0, A1 and A2 (see text) for galaxies in the group (blue histogram) compared to the full range of GZ classifications (grey histograms). Although not every machine learnt grouping can be described as a concordance group when compared to Galaxy Zoo classifications, this figure illustrates that the algorithm is creating groups that would have received a consistent human classification.

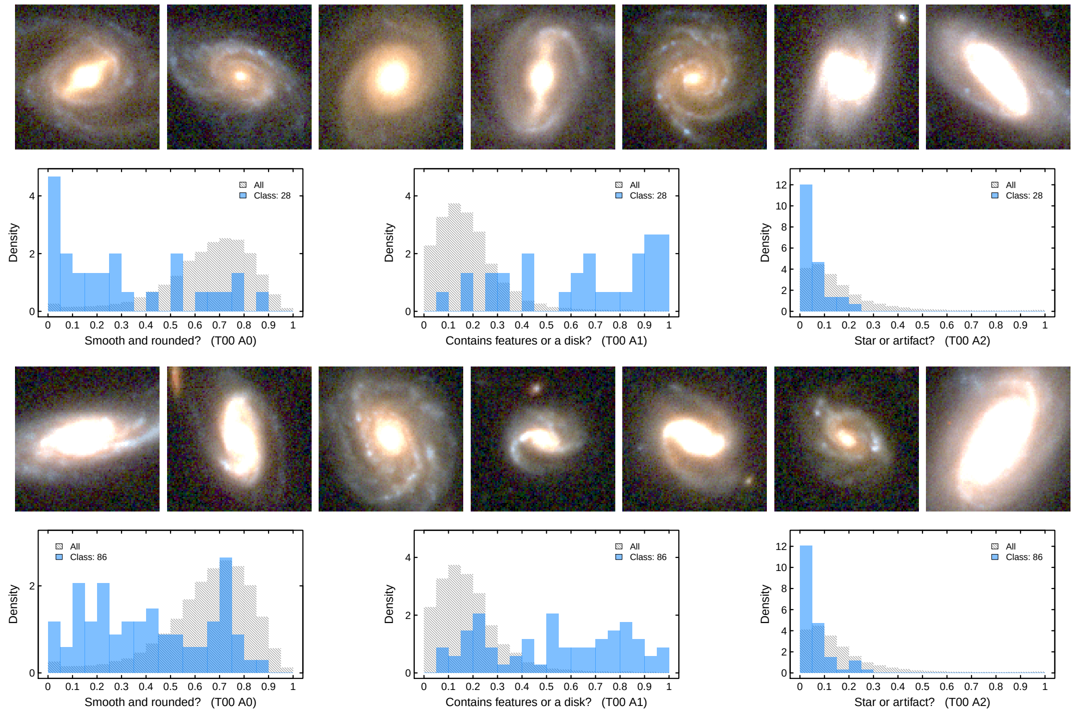


FIGURE 6.15: This figure shows two examples of what I refer to as a ‘concordance’ group, where over 50% of the galaxies for which Galaxy Zoo classifications were made have a weighted fraction over 0.5 for question T00 A1 ‘*does it contain features or a disk?*’. The images show the top seven matches in the group and the histograms compare the distribution of weighted fractions for questions T00 A0, A1 and A2 (see text) for galaxies in the group (blue histogram) compared to the full range of GZ classifications (grey histograms). Although not every machine learnt grouping can be described as a concordance group when compared to Galaxy Zoo classifications, this figure illustrates that the algorithm is creating groups that would have received a consistent human classification.

the fraction of machine learnt groups that contain each of the GZ clean classifications. One could argue that if a high fraction of clean classifications are contained within a small fraction of top level machine learnt groups, then the algorithm has successfully pigeon-holed the human classifications. On the other hand, these clean descriptors are rather broad, whereas even the coarsest level of machine learnt classification offers a way to segregate (for example) ‘smooth’ galaxies.

I consider each of the clean classifications described above and sort the list of top level machine learnt groups according to the number of galaxies matched to the clean lists. I then simply calculate the cumulative fraction of each clean list to assess the fraction of unique groups containing 50% and 100% of the clean classification galaxies. The results are given in Table 6.2, which lists the 50% and 100% fractions describing how the various clean classes are distributed within our machine learnt groups. For spiral, featured, clumpy and edge-on galaxies, the majority of the cleanly classified galaxies are contained within less than 10% of the top level groups. The

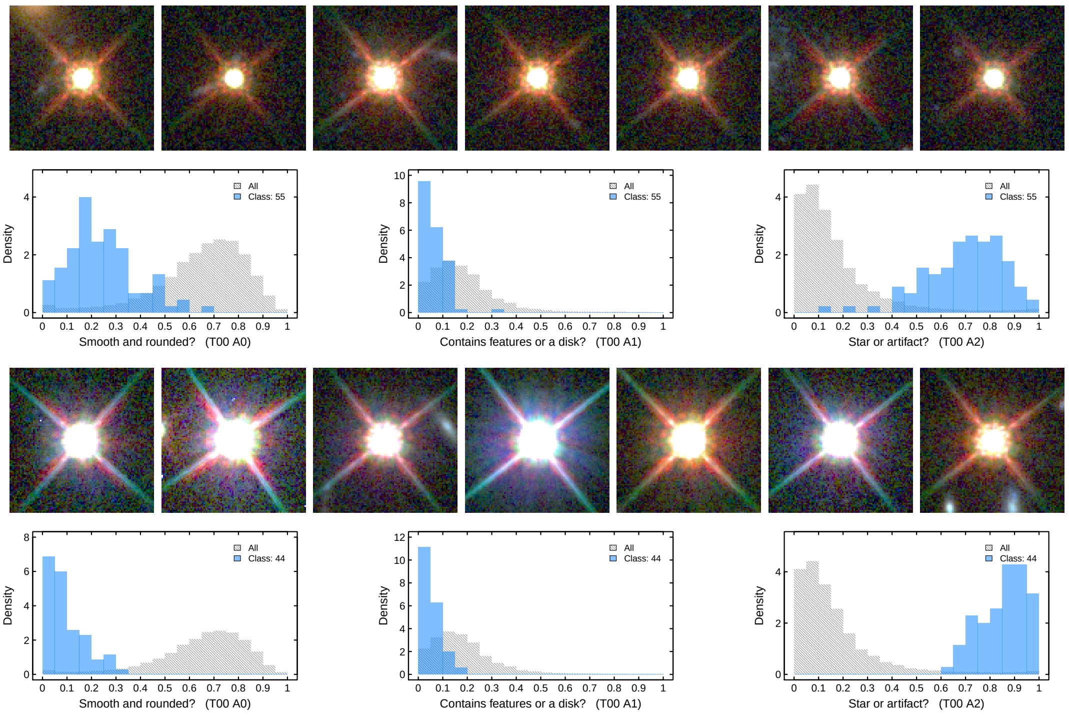


FIGURE 6.16: This figure shows two examples of what I refer to as a 'concordance' group, where over 50% of the galaxies for which Galaxy Zoo classifications were made have a weighted fraction over 0.5 for question T00 A2 '*is the target a star or artifact?*'. The images show the top seven matches in the group and the histograms compare the distribution of weighted fractions for questions T00 A0, A1 and A2 (see text) for galaxies in the group (blue histogram) compared to the full range of GZ classifications (grey histograms). Although not every machine learnt grouping can be described as a concordance group when compared to Galaxy Zoo classifications, this figure illustrates that the algorithm is creating groups that would have received a consistent human classification.

fraction is slightly higher for the smooth class. In all but the smooth class, 100% of the clean classifications are contained within around 50% of the machine learnt groups. For the smooth classification this is much higher – the galaxies seem to be spread over the majority of the machine learnt groups. This is perhaps unsurprising because the smooth classification dominates the clean class galaxies, and our algorithm has segregated these into a diverse set of sub-classes even at the top level of our hierarchical classification. Still, the fact that in all cases around half of the clean classifications are described by a minority of machine classes suggests that the algorithm is automatically classifying targets in a manner that is not dissimilar to a human inspector.

I conclude this section with a suggestion of an additional potential use for this technique which is to make predictions on which galaxies will be classified as, for example, clean_spiral by human classifiers. Indeed, blending the machine learning and human classification methods might be a particularly powerful technique; for instance, for extremely large samples of galaxies (or just large images), the algorithm could perform a 'first pass' unsupervised classification and feed subsamples of those results (blindly) to a cohort of human inspectors.

6.4 Summary

In this Chapter, I used the technique to analyse the *HST* Frontier Fields. By training the algorithm using galaxies from one field (Abell 2744) and applying the result to another (MACS0416.1-2403), I validated that the model generalised to new data. I also showed how the algorithm can cleanly separate early and late-type galaxies without any form of pre-directed training regarding what an 'early' or 'late' type galaxy is. I then applied the technique to the five *HST* CANDELS fields to create a catalogue of approximately 60000 classifications. I showed how the automatically identified groups of galaxies have similar morphological (and photometric) type. I have made the results public via a catalogue, a visual version of the catalogue and a galaxy similarity search. Finally, I compared the identified groups of galaxies to the human-classifications from the Galaxy Zoo: CANDELS project. Although there is not a direct mapping between Galaxy Zoo and our hierarchical labelling, I demonstrate a reasonable level of concordance between the galaxy zoo classifications and the groupings identified by the technique. Finally, I showed how the technique can be used to identify rarer objects and present lensed galaxy candidates from the CANDELS imaging.

In Chapter 1 I established a research gap in that a fully unsupervised machine learning technique applied to pixel data had not been demonstrated in astronomy. This Chapter presented the results of applying an entirely unsupervised machine learning technique to analyse the CANDELS fields using pixel data alone. I believe this is the first time that such a technique has successfully analysed a significant dataset in astronomy.

Chapter 7

Conclusions and Future Work

In this thesis I have presented an efficient unsupervised machine learning technique that uses a combination of GNG, hierarchical clustering and connected component labelling to explore astronomical surveys by automatically segmenting and labelling imaging data (as demonstrated in Chapter 4). The technique is a patch based model that does not process whole images of galaxies. Instead, it represents galaxies by combining many small overlapping patches. Each small overlapping patch is typically much smaller than the size of a galaxy, for example, a patch could contain a section of a spiral arm, or a section of a low surface brightness feature.

The development of the technique commenced by considering which image representations and unsupervised algorithms have the properties required to be effective. In Chapter 3 I identified three rotationally invariant image representations namely, the radially averaged power spectrum, RIFT and Spin Intensity. Rotational invariance is an essential characteristic enabling galaxies of similar type to be grouped together regardless of their orientation.

In the second half of Chapter 3 I considered the strengths and weaknesses of unsupervised algorithms of various types such as graph-based topological mapping, matrix factorisation and density modelling algorithms. After testing them on a high dimensional test dataset I decided to use the best performing algorithms GNG, Hierarchical Clustering and K-Means).

Having identified a selection of image representations and algorithms I then examined the available object detection and galaxy representation methods and how they could be applied to analyse astronomy survey image data. I decided to use a Connected-component labelling algorithm to identify which pixels belong to galaxies and to use a patch-based histogram representation of galaxies. The output of the Connected-component labelling algorithm is used to combine with the clustering of overlapping-patches to produce a histogram of each galaxy. This method is in contrast to standard methods such as using single images of galaxies. The decision to use histograms of small image patches was motivated by the need to identify and group galaxies with unusual shapes and sizes. It also enables galaxies that are very close to each other within the image to be analysed independently.

In Chapter 4 I described the model in its entirety and how it can segment survey images, locate and group similar objects. I presented the results of segmenting the Frontier Field survey images. However, I had not yet identified the optimum hyper-parameters of the model, such as which of three image representations was the most effective (from power spectrum, RIFT or Spin Intensity), and which algorithm combinations were best performing.

In Chapter 5 I described extensive tests designed to identify the optimal hyper-parameters, such as patch size, the best image representation and the best algorithm combinations. The technique is intended to be used for large surveys and so, unusually for an unsupervised machine learning mode, it was tested for generalisation. This involved analysing survey image data and then performing model selection by evaluating the results of identifying five types of galaxy on separate validation and test datasets. The data were from the *HST* Frontier Fields survey.

Once the best performing configurations of the model were identified it was necessary to evaluate the model using standard measures relevant to astronomers. The first test, described in the first sections of Chapter 6, was the demonstration that the model could cleanly separate early and late-type galaxies without any form of pre-directed training regarding what an ‘early’ or ‘late’ type galaxy is. The model analysed one Frontier Field (Abell 2744) for two types of galaxies and then successfully searched for the same types in the Frontier Field MACS0416.1-2403. In the second half of Chapter 6 I showed the results of applying the model to analyse the five *HST* CANDELS fields. I used the automatically identified groups to create my own catalogue of approximately 60000 CANDELS galaxies. I found that the galaxies in each of the groups do have similar morphological (and photometric) type. This is important evidence that my technique is identifying interesting groups of objects. Finally, I compared the identified groups of galaxies to the human-classifications from the Galaxy Zoo: CANDELS project. Although there is not a direct mapping between Galaxy Zoo and the catalogue, I demonstrated a reasonable level of concordance between human acquired labels and the groups of objects identified by the technique. Finally, I showed how the technique can be used to identify rarer objects and present lensed galaxy candidates from the CANDELS imaging.

7.1 Contribution

My major contribution corresponds to the gap I identified in Section 1.3 of Chapter 1, whereby existing work incorporating unsupervised machine learning algorithms to analyse astronomy images all use some form of supervision, such as the collation of a training dataset by pre-labelling galaxies. I established that a completely unsupervised machine learning technique that can be applied to explore imaging surveys without an upfront classification effort had not been proven.

This thesis has filled that gap by successfully establishing a completely unsupervised method to detect and then analyse galaxies in survey images with no upfront classification effort. My contributions are:

- I have established a novel technical framework using completely unsupervised machine learning methods that can identify and segment galaxies in survey images without a labelled training set.
- I have evaluated the proposed method and established that it can successfully identify similar galaxies and categorise them into groups. Also, the use of an overlapping-patch-based model is a novel technique previously unseen in astronomy. The unsupervised machine learning technique enables a galaxy similarity search that to my knowledge has not been demonstrated before. My resulting catalogue, a visual version of the catalogue and a galaxy similarity search are available at www.galaxyml.uk[1].
- The third contribution is the creation of a novel catalogue as a result of applying the technique to automatically analyse the five CANDELS fields (CANDELS is defined in Chapter 2). The CANDELS fields have been automatically analysed before[2]. For instance, Van der Wel et al. (2012) provided structural parameters for galaxies in CANDELS using the GALAPAGOS software (Barden et al., 2012) and Huertas-Company et al. (2015) produced a catalogue using supervised machine learning that classified galaxies into one of five types. The catalogue I produced, described in Chapter 6, is quite distinct from these other catalogues. It provides a much finer grouping of galaxies based on morphology and photometric characteristics.

[1] Source code is available at: https://github.com/alexhock/galaxymorphology

[2] I make the comparison with automatic computational methods. Therefore, I have not mentioned the classifications provided by the Galaxy Zoo:CANDELS project (Simmons et al., 2016a) which were created by crowd-sourcing.

One simple way of utilising my CANDELS classification catalogue is to use it to assemble samples of galaxies (or stars) that are photometrically and morphologically similar to a given test example. For example, one might have detailed observations of a specific galaxy in CANDELS and wish to find more examples of similar objects to build a statistical sample. One could simply match this target to the classification catalogue to find out which classification group it resides in, and therefore find all the other galaxies that 'look' (as far as the feature space allows) similar to it. Naturally, the selection function for this exercise would be complicated to understand (i.e. challenging to express in terms of, say, colour cuts), and that might be a limitation of this approach.

- The fourth contribution is the introduction and demonstration, for the first time, of the power spectrum feature as a successful image representation for unsupervised machine learning in this field. The power spectrum feature is defined in Chapter 3.

7.2 Future Work

7.2.1 Improving the Technique

7.2.1.1 Automatic Feature Representation

There are limitations to the method that should be noted. The most significant is the choice of the initial image representation. In this work we use sample vectors that effectively encode information about colour and intensity distribution on small (few pixels) scales. In principle the feature vector can be arbitrarily large, but at the cost of computation time; therefore there is a balance between performance and the sophistication of the chosen features. It is clear that the exact choice of image representation will have an impact on the ability of the algorithm to successfully segment and classify input data. It is possible that one could use an algorithm that identifies the optimal set of features to use (see unsupervised feature learning in Bengio et al. (2013), also stacked denoising autoencoders by Vincent et al. (2010)). Initial tests using convolutional autoencoders to learn features from the data can be seen in Figure 7.1. Careful attention to rotation invariance is required as the convolutional process is not invariant. Adopting the rotational invariance technique used in (Dieleman et al., 2015b) may be the solution.

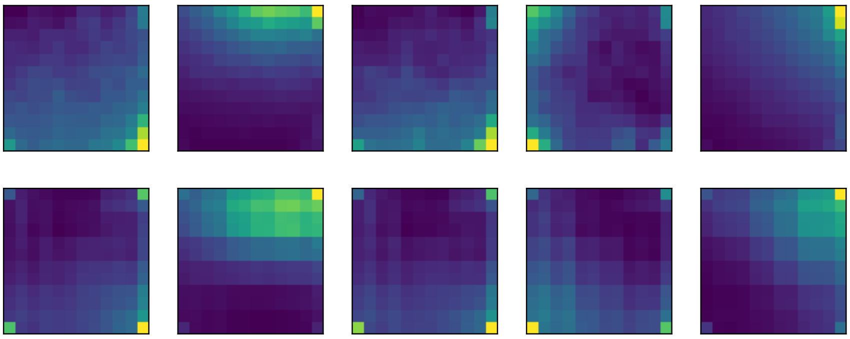


FIGURE 7.1: Autoencoded reconstructions. The top row contains original image patches and the bottom row shows the reconstructions created by the convolutional autoencoder. The reconstructions provide insight into how effectively the autoencoder has learnt the structure of the original image data.

7.2.1.2 Potential Improvements to Computational Performance

I have not fully optimized the algorithm for speed (and as noted above, performance will depend on the complexity of the image representations), however as a guide, the analysis of the Abell 2744 imaging took 36 msec per pixel and the analysis of the MACS 0416.1$-$2403 image with the model took 1.5 msec per pixel. The work was performed on a desktop computer with an Intel CPU. In my experience the amount of time spent on building the graph of the image patches takes the most time. This is why I've extensively tested for generalisation so that the graph can be re-used to analyse new images. These performances can clearly be dramatically improved, especially through the use of GPUs and optimal threading. The process is parallelisable making this a highly efficient algorithm to apply to large imaging data.

In addition, an approximate nearest neighbour algorithm (e.g. Muja and Lowe, 2014) could be used instead of the current approach in order to identify the patch type. This would be faster but potentially less accurate.

7.2.1.3 Improving Localisation

The technique currently makes no assumptions about object shape and size, but uses a simple threshold over the whole image to identify objects. However, the background level varies across the image and therefore a single threshold is not ideal. The introduction of a variable threshold that takes account of local background noise could lead to improved results. One approach would be to partition images and calculate a background level in each partition, however, galaxies may cross partition boundaries. The background map produced by Source Extractor (Bertin, 1996) could provide a useful example to follow.

7.2.1.4 Investigating Redshift Distributions

In Section 6.3 I discussed two possible explanations for individual clusters identifying galaxies at different redshifts. As discussed, this could be further explored by carrying out SED fitting to obtain physical parameters for galaxies in different groups. This would require multi-wavelength data for a large number of galaxies.

7.2.1.5 Investigating whether the high number of classification groups is a function of the method or a true representation of the underlying data

In Section 6.3 I discussed whether the high number of classification groups required in the CANDELS fields might be due to the underlying variation in the data (i.e. variation in colour and morphology), or whether it is a feature of the technique that a large number of groups are required to subdivide the parameter space. One method of testing this would be to create a number of dummy datasets with different degrees of variation and determine the optimum number of groups for each. If the number of groups required increases with increasing variation then this would confirm that the required number of groups depends on the underlying data.

7.2.2 Application to Future Surveys

The technique will be useful in the era of extremely large surveys such as the Large Synoptic Sky Telescope[3] (Ivezic et al., 2014, LSST) and *EUCLID* (Laureijs et al., 2011). The LSST is a ground-based 8.4 metre optical telescope designed to have a very wide field-of-view (3.5 degrees). This, and its capability to take a pair of images every 15 seconds, enable it to image the whole sky every three days. It has six wideband filters covering 350nm-1060nm. *EUCLID* is a space based telescope designed to observe the whole sky using optical (550 (green) to 920nm) and near-infrared (1000-2000nm) cameras.

[3] https://www.lsst.org

7.2.2.1 Unusual Objects

The unsupervised nature of the technique allows for the discovery of objects not previously known. As LSST and *EUCLID* are whole sky surveys they are expected to contain larger samples of unusual galaxies. I would expect the model to group these galaxies together. Alternatively, unusual galaxies may appear as outliers. These outliers can be found by searching at the fringe of the parameter space, for example, by looking for individual galaxies at the fringe of an identified group, or a group of galaxies that may itself be an outlier relative to other groups.

One type of known rare object that could be found in future surveys is strong galaxy lenses. Lenses magnify the flux and scale of background galaxies allowing analysis of galaxies of higher redshift than would be possible by direct imaging. The model developed in this thesis was successful in finding lenses in the CANDELS fields (see Figure 6.13 in Chapter 6). The *HST* CANDELS data has higher resolution than the LSST as the LSST is subject to the 'seeing' conditions provided by the atmosphere. Whereas the lensing features found in CANDELS images can be clearly seen (once they have been found), similar lensing features in LSST imaging are likely to appear as blue blobs or smudges. It may be possible to change the CANDELS data to test whether the technique could find similar lenses in LSST data by degrading the resolution of CANDELS data and then re-running the method to see if they are still grouped together. However, this should not be a problem for *EUCLID* images.

7.2.2.2 Transients

One of the goals of the LSST is to 'make a high-definition colour movie of the deep Universe' (Ivezic et al., 2014). By imaging the sky every few nights transient objects that appear and disappear over short time scales can be detected. The LSST data pipeline will provide small cutouts around potential detected transient objects. It is expected that many of the detections will be false positives[4]. The technique could be adapted to analyse these potential transient detections to confirm whether they are real transients, such as supernovae, or false detections due to systematics.

The technique could be applied by collating all the transients for a period of time and applying the model to these data only. The technique would classify similar transients together, creating a series of groups of false positives and a series of groups of actual transients. Once enough

[4] https://www.lsst.org/about/dm/petascale

transients have been acquired, the identified groups could be used to classify new transients in real time.

7.2.2.3 Application to Radio Surveys

The Square Kilometre Array is anticipated to produce Petabytes of radio data that will require automated analysis[5]. The technique can not easily be applied to radio intensity images as they contain so little morphology. However, it may be possible to apply the technique to combined optical and radio data.

7.3 Final Summary

I have established a completely unsupervised machine learning technique that can segment, locate and group similar objects in image surveys for the first time. Unlike previous techniques, it requires no training set or pre-labelling of galaxies. I have identified the hyper-parameters, image representations and algorithms that show the highest performance, and I have applied the model to produce a novel catalogue of the five *HST* CANDELS fields. The results of this catalogue can be seen at www.galaxyml.uk.

[5] www.skatelescope.org

Appendix A

Calculating Image Features

A.1 Calculating Image Gradients

The Histogram of Oriented Gradients (HoG) and Rotationally Invariant Feature Transform (RIFT) features described in Chapter 3 use image gradients to create a histogram representation of an image or image patch. There are many ways to approximate the image gradient using the convolutional operation. Several standard kernels such as the Sobel operator (Sobel, 1990). The popular method presented by Dalal and Triggs (2005) calculates image gradients in the following way:

1. Calculate the image gradient in the horizontal and vertical directions by applying a kernel for each direction $\begin{bmatrix}-1,0,1\end{bmatrix}$ and $\begin{bmatrix}-1\\0\\1\end{bmatrix}$. A 2×1 kernel could be used, however, it will shift the image by half a pixel. To avoid this 3×1 kernels are used. The result is two output images, gx and gy ,one highlighting the gradient changes in the x direction and the second highlighting the gradient changes in y direction. Figure A.1 shows the effect of this process for a test image and a real image.

2. Use the two output images from the first step to calculate the magnitudes of the gradients using $\sqrt{gx_{xy}^2 + gy_{xy}^2}$ and the orientations of the gradients using $atan(gy_{xy}/gx_{xy})$ where xy are the pixel positions.

3. Create a histogram or orientations by binning into a set number of bins representing angular directions. A typical number of bins is eight with each bin representing $360°/8 = 45°$.

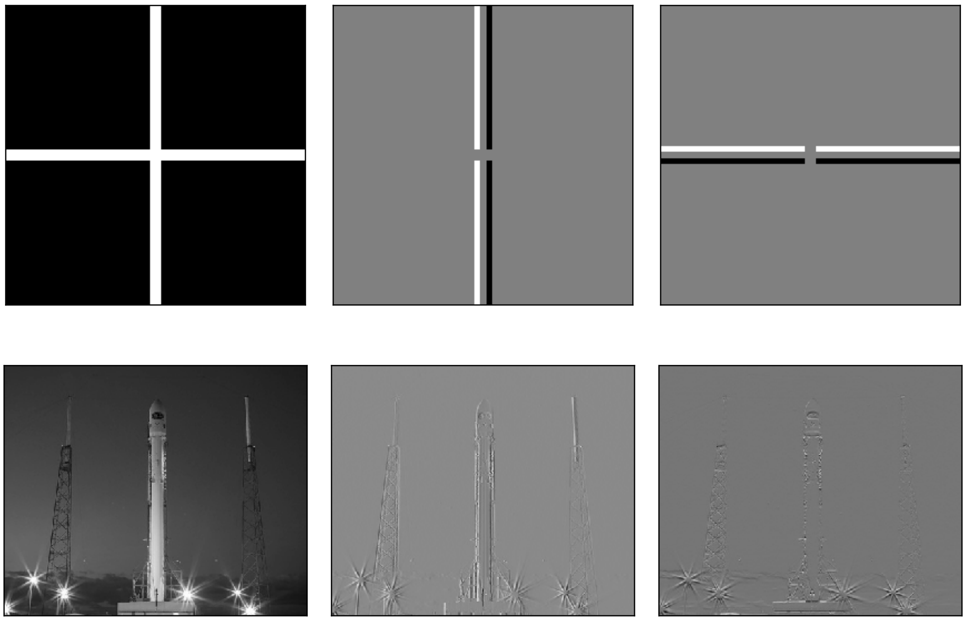


FIGURE A.1: The top left image is a toy image with zero values at every location except for the vertical and horizontal lines with the value 1. The centre image is the output of applying the horizontal kernel. We can see the detection of the increase in gradient, the white vertical line, and the decrease in gradient, the black vertical line. The image to the top right shows the result of applying the vertical kernel. The three images in the bottom row show the effect on a real image of a rocket (left image). We can see that the vertical features are highlighted in the middle images, and the horizontal features are highlighted in the right images.

The value that is added to the bin is based on the magnitude of the gradient. Figure 3.1 shows a visual depiction of HoG histograms at each cell location.

The key difference between HoG and RIFT is the regions of the gradient image used to create the histograms in the final step. HoG uses ‘cells’ which are square windows of a fixed size and a histogram is calculated for each cell in the image. The final feature vector is obtained by concatenating all histograms. HoG is not rotationally invariant feature so RIFT uses circular annular bins around a centre point and the histograms are created by binning the pixel gradients in these annular bins.

A.2 Scale Space

Multiple scales exist in an image, for example, in an image of a city we have no problem identifying the large scale structures such as the buildings, even though they consist of many small features such as bricks, door and windows. We also have no problem identifying objects that are nearer to the camera that maybe smaller such as people in front of building or the cracks in a pavement or road in front of a building. To us, the details of an image disappear as we move away from an object because our attention is drawn to the larger scales. Computer vision researchers have attempted to incorporate this concept called scale space into their models since the 90s (Lindeberg, 1994). The basic idea of scale-space theory is to consider descriptions at multiple scales in order to be able to capture the unknown scale variations that may occur. The main type of scale space is the linear, Gaussian, scale space. Where scale space representation is a family of derived signals defined by the convolution of source image with the two dimension Gaussian kernel. The variance parameter of the Gaussian kernel is called the scale parameter. As the value of the scale parameter increases the scale space representation becomes smoother and smoother. Which means more and more of the original image details are removed. An example is shown in Figure A.2. Although the work on representing image structures at multiple scales has been well established, in many cases, we still need to select locally appropriate scales for further analysis or study.

We note that the concept of scale space in computer vision has largely been superseded by using multi-layer neural networks (convolutional nets) for object detection and image segmentation Ren et al. (2015); Redmon et al. (2016); Long et al. (2015).

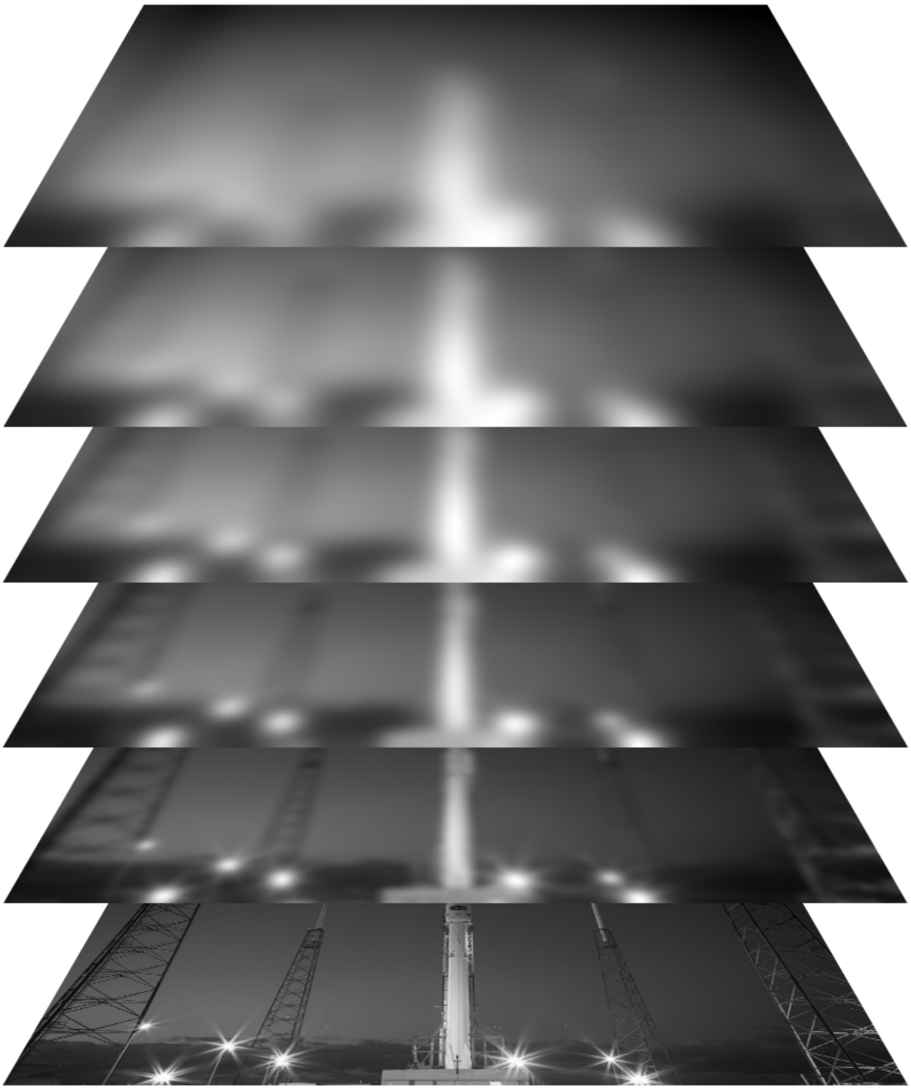


FIGURE A.2: An example representation of the scale space of an image. The image at the bottom is a standard grayscale image of a rocket. The images above it have an increasingly large Gaussian blur applied which removes detail until only the largest structures remain in the top image.

A.3 Applying the Discrete Fourier Transform to Images

A signal can be approximated using a sum of individual sinusoidal frequency components. These individual frequency components can be identified using an algorithm called the Discrete Fourier Transform (Smith et al., 1997, DFT).

The DFT, defined in equation A.1 and with Euler's identity in A.2, converts a sequence of N complex numbers $x_0, x_1, ..., x_{N-1}$ to a new sequence of N complex numbers, for $0 \leq k \leq N-1$. The x_i are the signal values at equally spaced times $t = 0, 1, ..., N-1$. The output X_k is a complex number encoding the amplitude and the phase of a sinusoidal wave with frequency k/N cycles per unit time. The goal, when applying the DFT, is to find the coefficients X_k that describe the sinusoid components which approximate the signal.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \tag{A.1}$$

$$X_k = \sum_{n=0}^{N-1} x_n \cdot [cos(2\pi kn/N) - i \cdot sin(2\pi kn/N)] \tag{A.2}$$

The time complexity of the DFT is $O(n^2)$. Famously Cooley and Tukey (1965) introduced a Fast Fourier Transform (FFT) algorithm capable of computing the coefficients in $O(N \log N)$ although it appears that Gauss was perhaps the first person to identify such an algorithm (Heideman et al., 1984).

A.3.1 1D Power Spectrum

The Power Spectral Density (PSD), commonly known as the power spectrum, is the signal power as a function of frequency. It identifies the individual frequencies that contribute to the power of the signal. It is calculated by taking the square of the Fourier transform's magnitude. Figure A.3 calculates the power spectrum for two example time series.

$$Magnitude\ of\ Power = \frac{|DFT(signal)|^2}{N} \tag{A.3}$$

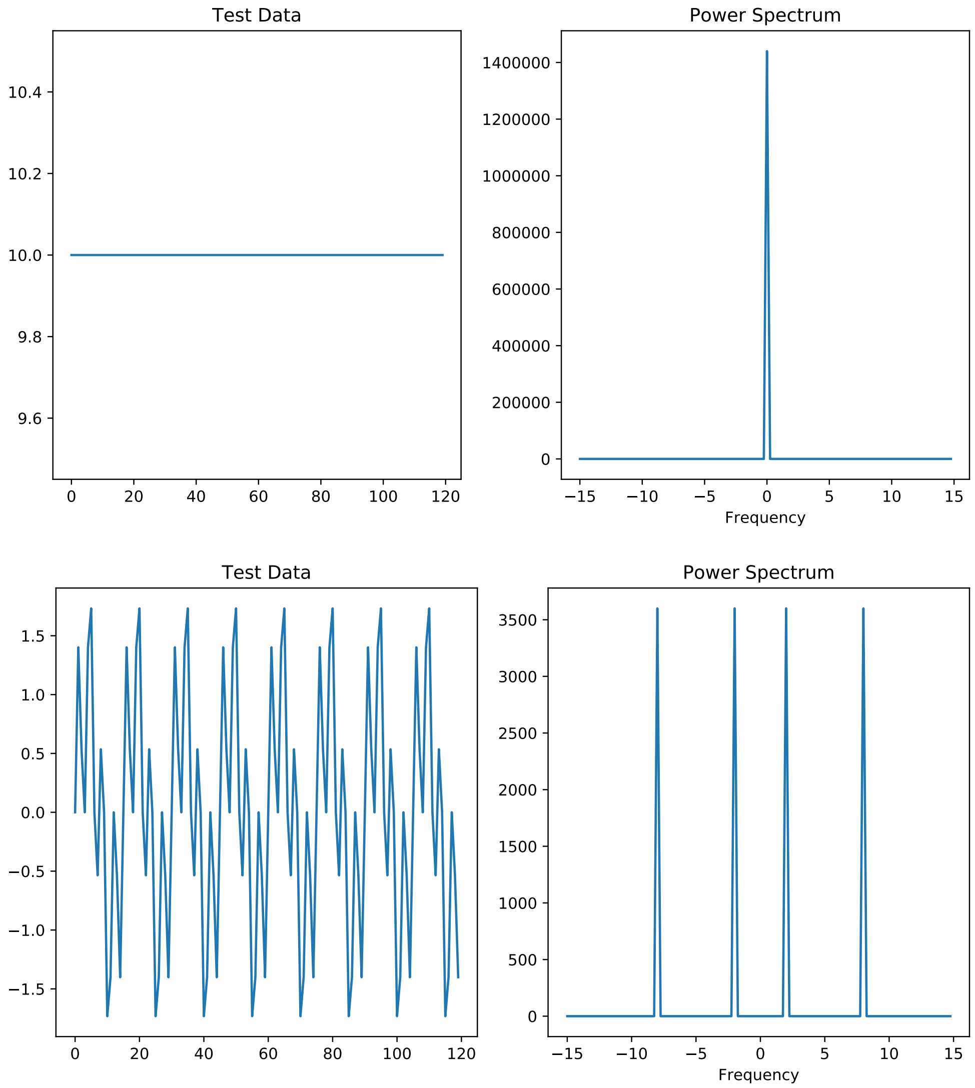


FIGURE A.3: The power spectrum of a time series. The power spectrum is symmetrical as can be seen in the right hand graphs. The zero frequency represents the average across the time series. For the power spectrum at the top the average is 10. The power spectrum shows a corresponding peak for the zero frequency. The lower two graphs show on the left test data consisting of two sine waves, one with a frequency of 2 and the second with a frequency of 8. Note that the average of the time series is 0. The corresponding power spectrum on the right shows the peaks at 2 and 8. The power spectrum is symmetrical.

When applied to images which have a finite spatial size (the equivalent of a finite time window) we can use the energy spectral density which is calculated by multiplying the DFT by it's complex conjugate.

A.3.2 2D Power Spectrum

The DFT is not limited to one dimensional time series data; it can also be used to calculate the Fourier co-efficients for data with n dimensions. In image processing we can use a 2D DFT to calculate the transform coefficients for the two dimensional pixel values in an image. The frequencies are the changes in pixel intensity across the image, also known as spatial frequency of pixel intensity variation as in Figure A.4. The steps to calculate the 2D DFT of an image are:

1. Calculate the 1D DFT for each row of pixels in the image. Each pixel intensity value is a discrete sample. The DFT calculation when applied to this data results in a matrix of complex numbers with the same dimensions as the input image. Each row of the result contains the complex coefficients of the DFT for each row of the original image. Note that the zero coefficient is the average brightness of each row of the original image. Also the results are symmetrical across the middle column i.e. the right hand side is the reverse of the left hand side.

2. The next step is to calculate the DFT of each column of the result from the first step i.e. treat each column of values in the co-efficient matrix as a signal and compute its DFT coefficients. The result is another 2D matrix of the same size containing the completed 2D Fourier transform of the original image.

In practice we do not follow these steps, we use a 2D FFT implementation of standard software libraries.

Once the 2D DFT is calculated we can then calculate the 2D power spectrum for the image by using the same power spectrum equations as before i.e. computing the absolute square of the complex numbers in each of the the elements of the 2D DFT matrix. (Note the absolute square of a complex number is always real). The 2D power spectrum is also symmetrical.

To provide an intuition of what the power spectrum is doing Figure A.5 shows the 2D power spectrum applied to three images each with a 2D sine wave with a different frequency for each image. These images were created by copying a discrete time series signal and repeating the signal to every row in the image. Figure A.6 shows the effect on the power spectrum of rotating the original image. We can see that the power spectrum changes when the image is rotated. In computer vision terminology, it is not invariant to rotation. Therefore in Chapter 3 we add the averaging over annular bins, which results in a rotationally invariant representation.

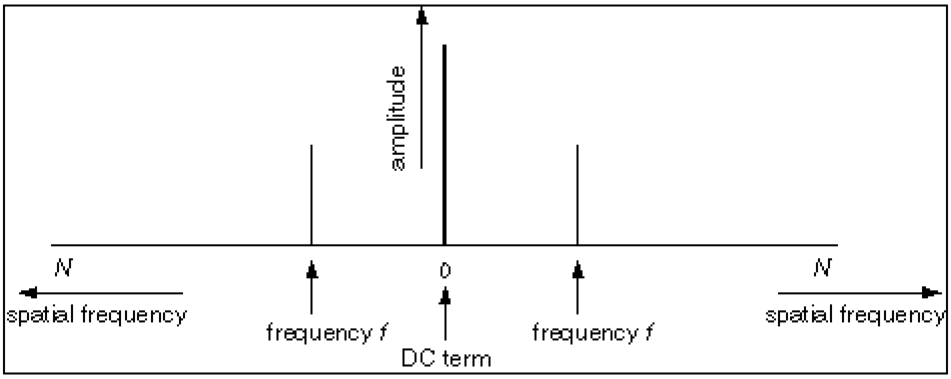


FIGURE A.4: When calculating a 2D power spectrum the frequency represents spatial frequency or how the pixel intensities vary across the image. The ”DC term” corresponding to zero frequency represents the average brightness across the image.

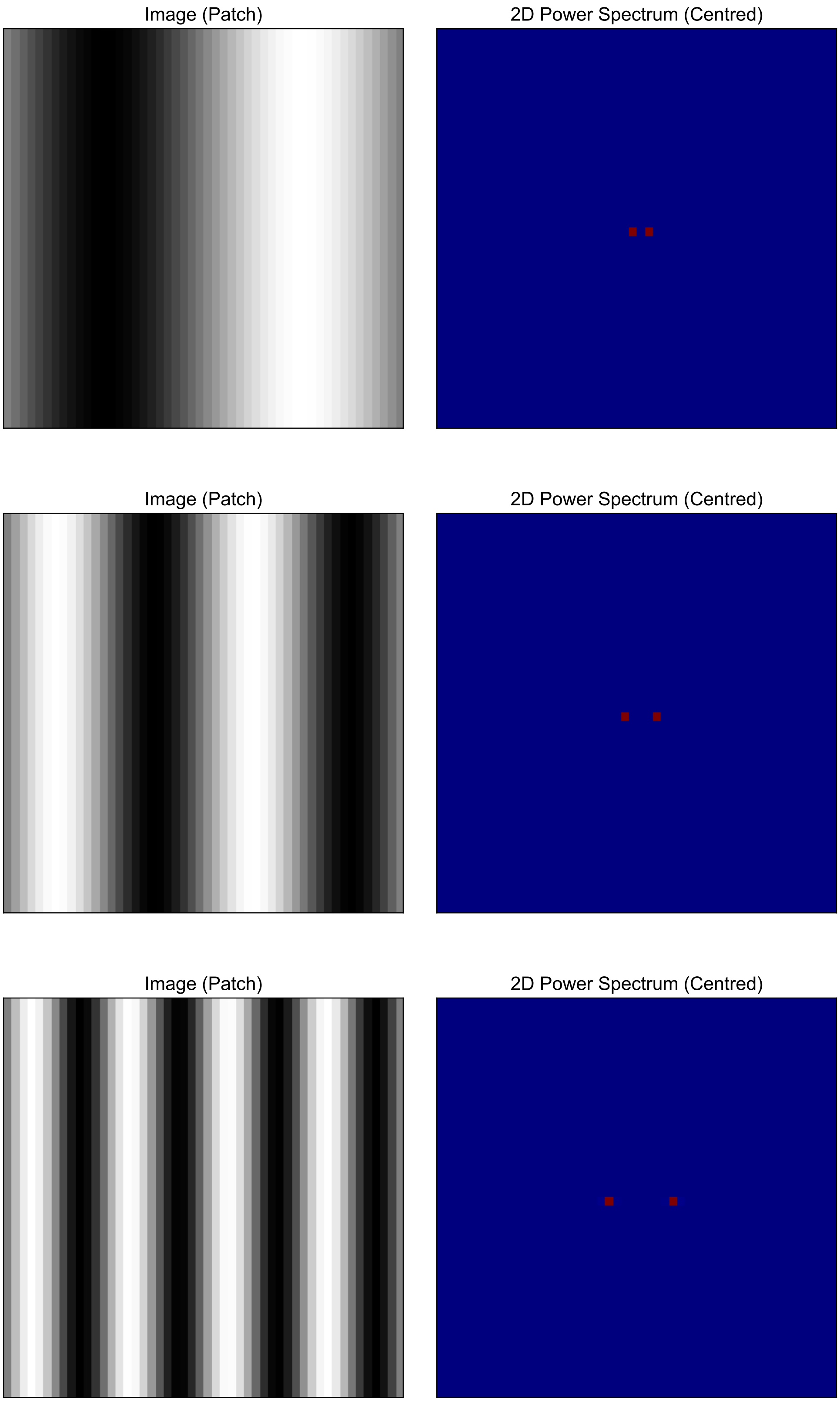


Image (Patch)

2D Power Spectrum (Centred)

Image (Patch)

2D Power Spectrum (Centred)

Image (Patch)

2D Power Spectrum (Centred)

FIGURE A.5: The images to the left have a sine wave with spatial frequency of 1 (top), 2 (middle) and 4 (bottom). The right three images show the corresponding 2D power spectra. The higher the spatial frequency the further from the central zero frequency. A smooth image has high power values in the centre, and edges or textures with large changes in pixel variation over short distances appear as larger values away from the centre.

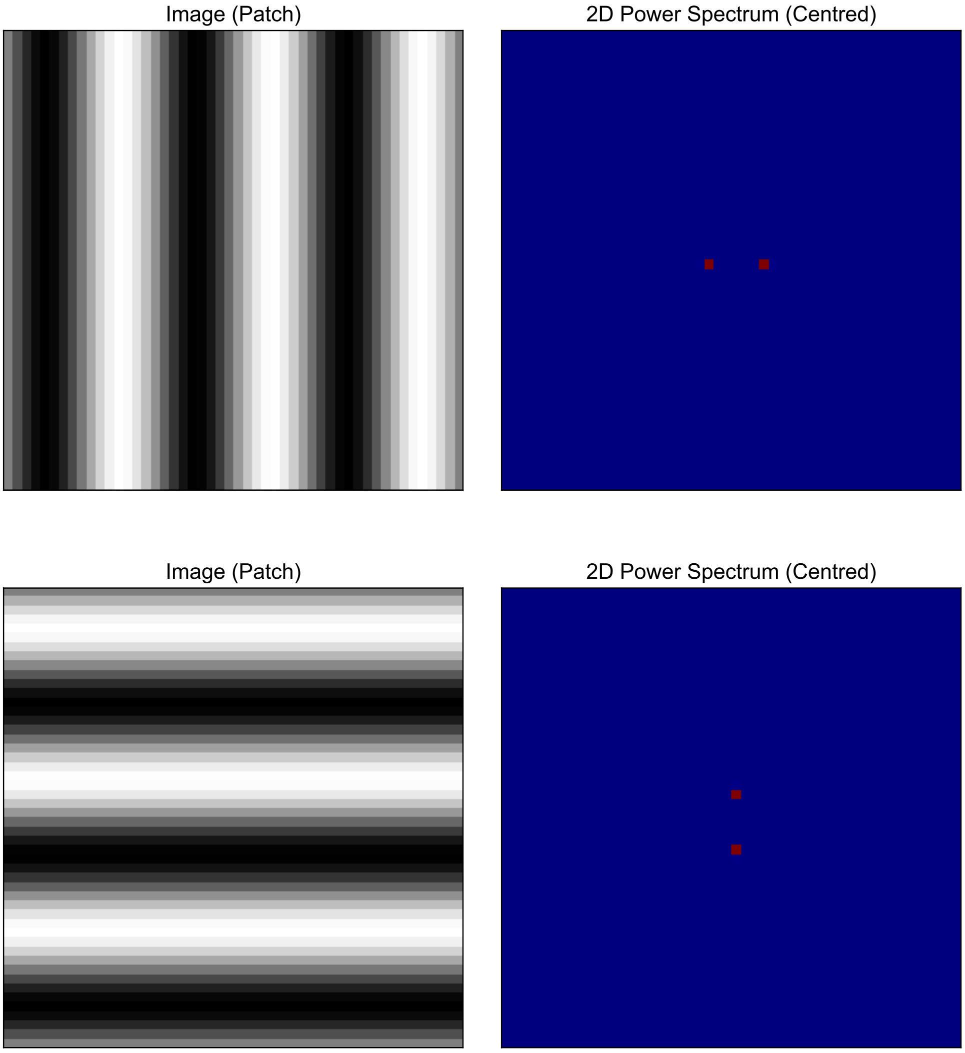


FIGURE A.6: The image to the top left contains a spatial frequency of three. The image to the bottom left is the same image rotated 90°. The top right image shows the 2D power spectrum of the top left image, and the bottom right image shows the 2D power spectrum of the bottom left image. This shows the effect on the 2D power spectrum of rotations. The power spectrum has the same 90° rotation.

Appendix B

Performance Evaluation

B.1 Adjusted Rand Index

Adjusted Rand Index (ARI) proposed by Hubert and Arabie (1985) is recommended as the index of choice for measuring agreement between two partitions in a clustering analysis with different numbers of clusters (Santos and Embrechts, 2009). Since it was originally introduced the ARI has become one of the most successful cluster evaluation measures (Taşdemir et al., 2015; Kulesza et al., 2014; Frank et al., 2014).

To compute the ARI a contingency table is created for two clusterings. The table is created by identifying how pairs of data points are matched. The ARI represents an adjusted frequency of occurrence of agreements over the total pairs. I now show how to calculate ARI using a simple example. I start with a small dataset consisting of 12 data points. The dataset has been clustered using two algorithms K-Means and Gaussian Mixture Models (GMM) where both used a pre-set number of clusters, k, set to 3. To calculate ARI all we need are the two sets of cluster assignments of the 12 data points, the details of the clustering algorithms are not required.

The equation for calculating ARI is[1]:

$$ARI = \frac{A - \frac{B}{C}}{\frac{D}{2} - \frac{B}{C}}$$

where $A = \sum_{i,j} \binom{n_{i,j}}{2}$, $B = \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}$, $C = \sum_j \binom{b_j}{2}$ and $D = \sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}$

[1] source: https://wiki.cs.byu.edu/cs-401r/adjusted-rand-index

The contingency table shows the counts of the co-occurrences between the two clusterings of the 12 data points. The table shows that GMM clustered the 12 data points into the 3 clusters evenly, with 4 data points in each cluster. K-Means clustered the 12 data points unevenly with 4 datapoints in cluster 1, 5 in cluster 2, and 3 data points in cluster 3:

	GMM Cluster 1	GMM Cluster 2	GMM Cluster 3	Row Sums
K-Means Cluster 1	$n_{1,1} = 1$	$n_{1,2} = 0$	$n_{1,3} = 3$	$a_1 = 4$
K-Means Cluster 2	$n_{2,1} = 2$	$n_{2,2} = 2$	$n_{2,3} = 1$	$a_2 = 5$
K-Means Cluster 3	$n_{3,1} = 1$	$n_{3,2} = 2$	$n_{3,3} = 0$	$a_3 = 3$
Column Sums	$b_1 = 4$	$b_2 = 4$	$b_3 = 4$	12

The A, B, C and D parameters are calculated from the contents of the contingency table:

$$A = \sum_{i,j} \binom{n_{i,j}}{2} = 2\binom{0}{2} + 3\binom{1}{2} + 3\binom{2}{2} + 1\binom{3}{2} = 0 + 0 + 3 + 3 = 6$$

$$B = \sum_{i,j} \binom{a_i}{2} \sum_j \binom{b_j}{2} = \left(\binom{4}{2} + \binom{5}{2} + \binom{3}{2}\right) \times 3\binom{4}{2} = (6 + 10 + 3) \times (3 \times 6) = 19 \times 18 = 342$$

$$C = \binom{n}{2} = \binom{12}{2} = 66$$

$$D = \sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} = 19 + 18 = 37$$

$$ARI = \frac{6 - \frac{342}{66}}{\frac{37}{2} - \frac{342}{66}} = 0.0614$$

Bibliography

Abazajian, K.N., Adelman-McCarthy, J.K., Agüeros, M.A., et al., 2009. The seventh data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 182(2):543.

Abraham, R.G., Van Den Bergh, S., and Nair, P., 2003. A new approach to galaxy morphology. i. analysis of the sloan digital sky survey early data release. *The Astrophysical Journal*, 588(1):218.

Alcantarilla, P.F., 2011. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7):1281.

Alcantarilla, P.F., Bartoli, A., and Davison, A.J., 2012. Kaze features. In *European Conference on Computer Vision*, pages 214–227. Springer.

Allen, P.D., Driver, S.P., Graham, A.W., et al., 2006. The millennium galaxy catalogue: bulge–disc decomposition of 10 095 nearby galaxies. *Monthly Notices of the Royal Astronomical Society*, 371(1):2.

Aniyan, A. and Thorat, K., 2017. Classifying radio galaxies with the convolutional neural network. *The Astrophysical Journal Supplement Series*, 230(20):15pp.

Ascaso, B., Wittman, D., and Benítez, N., 2012. Bayesian cluster finder: clusters in the cfhtls archive research survey. *Monthly Notices of the Royal Astronomical Society*, 420(2):1167.

Bachem, O., Lucic, M., Hassani, H., et al., 2016. Fast and provably good seedings for k-means. In *Advances in Neural Information Processing Systems*, pages 55–63.

Baillard, A., Bertin, E., De Lapparent, V., et al., 2011. The efigi catalogue of 4458 nearby galaxies with detailed morphology. *Astronomy & Astrophysics*, 532:A74.

Baldry, I.K., Glazebrook, K., Brinkmann, J., et al., 2004. Quantifying the bimodal color-magnitude distribution of galaxies. *The Astrophysical Journal*, 600(2):681.

Ball, N.M., Loveday, J., Fukugita, M., et al., 2004. Galaxy types in the sloan digital sky survey using supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 348(3):1038.

Banerji, M., Lahav, O., Lintott, C.J., et al., 2010. Galaxy Zoo: reproducing galaxy morphologies via machine learning. *MNRAS*, 406:342.

Barden, M., Häußler, B., Peng, C.Y., et al., 2012. Galapagos: from pixels to parameters. *Monthly Notices of the Royal Astronomical Society*, 422(1):449.

Baron, D. and Poznanski, D., 2016. The weirdest sdss galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society*, 465(4):4530.

Bay, H., Ess, A., Tuytelaars, T., et al., 2008. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346.

Beck, M.R., Scarlata, C., Fortson, L.F., et al., 2018. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society*, 476(4):5516.

Bengio, Y., Courville, A., and Vincent, P., 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798.

Bertin, E., 2012. Displaying Digital Deep Sky Images. In P. Ballester, D. Egret, and N.P.F. Lorente, editors, *Astronomical Data Analysis Software and Systems XXI*, volume 461 of *Astronomical Society of the Pacific Conference Series*, page 263.

Bertin, E.; Arnouts, S., 1996. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement*, 117:393.

Birant, D. and Kut, A., 2007. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208.

Bishop, C.M., 2006. *Pattern recognition and machine learning*. springer.

Blanton, M.R., Bershady, M.A., Abolfathi, B., et al., 2017. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. *AJ*, 154:28.

Bonfield, D.G., Sun, Y., Davey, N., et al., 2010. Photometric redshift estimation using Gaussian processes. *MNRAS*, 405:987.

Brammer, G.B., van Dokkum, P.G., and Coppi, P., 2008. Eazy: A fast, public photometric redshift code. *The Astrophysical Journal*, 686(2):1503.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1):5.

Brescia, M., Cavuoti, S., D'Abrusco, R., et al., 2013. Photometric Redshifts for Quasars in Multi-band Surveys. *ApJ*, 772:140.

Bullock, J., 2012. Hubble deep fields initiative 2012 science working group report.

Caon, N., Capaccioli, M., and D'onofrio, M., 1993. On the shape of the light profiles of early-type galaxies. *Monthly Notices of the Royal Astronomical Society*, 265(4):1013.

Cardamone, C., Schawinski, K., Sarzi, M., et al., 2009. Galaxy zoo green peas: discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, 399(3):1191.

Carrasco Kind, M. and Brunner, R.J., 2014. SOMz: photometric redshift PDFs with self-organizing maps and random atlas. *MNRAS*, 438:3409.

Cavuoti, S., Brescia, M., Longo, G., et al., 2012. Photometric redshifts with the quasi Newton algorithm (MLPQNA) Results in the PHAT1 contest. *A&A*, 546:A13.

Chang, F., Chen, C.J., and Lu, C.J., 2004. A linear-time component-labeling algorithm using contour tracing technique. *computer vision and image understanding*, 93(2):206.

Charnock, T. and Moss, A., 2017. Deep recurrent neural networks for supernovae classification. *The Astrophysical Journal Letters*, 837(2):L28.

Chen, X., Duan, Y., Houthooft, R., et al., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180.

Coates, A., Ng, A., and Lee, H., 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223.

Collister, A.A. and Lahav, O., 2004. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *PASP*, 116:345.

Conselice, C.J., 2003. The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1.

Conselice, C.J., 2014. The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy and Astrophysics*, 52:291.

Cooley, J.W. and Tukey, J.W., 1965. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297.

D'Abrusco, R., Fabbiano, G., Djorgovski, G., et al., 2012. CLaSPS: A New Methodology for Knowledge Extraction from Complex Astronomical Data Sets. *ApJ*, 755:92.

Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Dalal, N., Triggs, B., and Schmid, C., 2006. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer.

Davis, D.R. and Hayes, W.B., 2014. SpArcFiRe: Scalable Automated Detection of Spiral Galaxy Arm Segments. *ApJ*, 790:87.

de Vaucouleurs, G., 1948. Recherches sur les nebuleuses extragalactiques. In *Annales d'Astrophysique*, volume 11, page 247.

De Vaucouleurs, G., 1959. Classification and morphology of external galaxies. In *Astrophysik iv: Sternsysteme/astrophysics iv: Stellar systems*, pages 275–310. Springer.

De Vaucouleurs, G., 1964. The luminosity classification of galaxies and some applications. Technical report, MCDONALD OBSERVATORY AUSTIN TEX.

Déniz, O., Bueno, G., Salido, J., et al., 2011. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598.

Dieleman, S., Willett, K.W., and Dambre, J., 2015a. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441.

Dieleman, S., Willett, K.W., and Dambre, J., 2015b. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441.

Djorgovski, S.G., Mahabal, A., Drake, A., et al., 2013. *Sky Surveys*, page 223.

Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

Einstein, A., 1936. Lens-like action of a star by the deviation of light in the gravitational field. *Science*, 84(2188):506.

Ellis, R.S., 2010. Gravitational lensing: a unique probe of dark matter and dark energy. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1914):967.

Ester, M., Kriegel, H.P., Sander, J., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Faure, C., Anguita, T., Alloin, D., et al., 2011. On the evolution of environmental and mass properties of strong lens galaxies in COSMOS. *A&A*, 529:A72.

Figueiredo, M.A.T. and Jain, A.K., 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381.

Fiorio, C. and Gustedt, J., 1996. Two linear time union-find strategies for image processing. *Theoretical Computer Science*, 154(2):165.

Firth, A.E., Lahav, O., and Somerville, R.S., 2003. Estimating photometric redshifts with artificial neural networks. *MNRAS*, 339:1195.

Fiser, D., Faigl, J., and Kulich, M., 2012. Growing neural gas efficiently. *Neurocomputing*.

Frank, C., Land, W.M., Popp, C., et al., 2014. Mental representation and mental practice: experimental investigation on the functional links between motor memory and motor imagery. *PloS one*, 9(4):e95175.

Freeman, K.C., 1970. On the disks of spiral and s0 galaxies. *The Astrophysical Journal*, 160:811.

Freeman, P.E., Izbicki, R., Lee, A.B., et al., 2013. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282.

Fritzke, B., 1995. A Growing Neural Gas Network Learns Topologies. *Advances in Neural Information Processing Systems 7*, 7.

Fritzke, B. et al., 1995. A growing neural gas network learns topologies. *Advances in neural information processing systems*, 7:625.

Fustes, D., Manteiga, M., Dafonte, C., et al., 2013. An approach to the analysis of SDSS spectroscopic outliers based on Self-Organizing Maps. *ArXiv e-prints*.

Geach, J.E., 2012. Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys. *MNRAS*, 419:2633.

Gers, F.A., Schmidhuber, J., and Cummins, F., 1999. Learning to forget: Continual prediction with lstm.

Gini, C., 1912. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.*

González-López, J., Bauer, F., Romero-Cañizales, C., et al., 2017. The alma frontier fields survey-i. 1.1 mm continuum detections in abell 2744, macs j0416. 1-2403 and macs j1149. 5+ 2223. *Astronomy & Astrophysics*, 597:A41.

Goodfellow, I., Bengio, Y., and Courville, A., 2016. *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goodfellow, I.J., Warde-Farley, D., Mirza, M., et al., 2013. Maxout networks. *arXiv preprint arXiv:1302.4389*.

Graham, A.W., 2013. Elliptical and disk galaxy structure and modern scaling laws. In *Planets, Stars and Stellar Systems*, pages 91–139. Springer.

Graves, A., Mohamed, A.r., and Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Grogin, N.A., Kocevski, D.D., Faber, S., et al., 2011. Candels: the cosmic assembly near-infrared deep extragalactic legacy survey. *The Astrophysical Journal Supplement Series*, 197(2):35.

Hart, R.E., Bamford, S.P., Hayes, W.B., et al., 2017. Galaxy zoo and sparcfire: Constraints on spiral arm formation mechanisms from spiral arm number and pitch angles. *Monthly Notices of the Royal Astronomical Society*, 472(2):2263.

Hastie, T., Tibshirani, R., and J, F., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Häußler, B., Bamford, S.P., Vika, M., et al., 2013. Megamorph–multiwavelength measurement of galaxy structure: complete sérsic profile information from modern surveys. *Monthly Notices of the Royal Astronomical Society*, 430(1):330.

He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, L., Chao, Y., and Suzuki, K., 2008. A run-based two-scan labeling algorithm. *Image Processing, IEEE Transactions on*, 17(5):749.

Heideman, M., Johnson, D., and Burrus, C., 1984. Gauss and the history of the fast fourier transform. *IEEE ASSP Magazine*, 1(4):14.

Hocking, A., Sun, Y., Geach, J., et al., 2017. Mining hubble space telescope images. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE.

Holmberg, E., 1958. A photographic photometry of extragalactic nebulae. *Meddelanden fran Lunds Astronomiska Observatorium Serie II*, 136:1.

Hoshen, J. and Kopelman, R., 1976. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Phys. Rev. B*, 14:3438.

Hubble, E., 1926. No. 324. extra-galactic nebulae. *Contributions from the Mount Wilson Observatory/Carnegie Institution of Washington*, 324:1.

Hubble, E.P., 1936. *The realm of the nebulae*, volume 25. Yale University Press.

Hubert, L. and Arabie, P., 1985. Comparing partitions. *Journal of classification*, 2(1):193.

Huertas-Company, M., Aguerri, J.A.L., Bernardi, M., et al., 2011. Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification. *A&A*, 525:A157.

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al., 2015. A catalog of visual-like morphologies in the 5 candels fields using deep learning. *The Astrophysical Journal Supplement Series*, 221(1):8.

Huertas-Company, M., Rouan, D., Tasca, L., et al., 2008. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images-i. method description. *Astronomy & Astrophysics*, 478(3):971.

Huertas-Company, M., Tasca, L., Rouan, D., et al., 2009. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. II. Quantifying morphological k-correction in the COSMOS field at 1 ¡ z ¡ 2: Ks band vs. I band. *A&A*, 497:743.

in der Au, A., Meusinger, H., Schalldach, P.F., et al., 2012. ASPECT: A spectra clustering tool for exploration of large spectral surveys. *A&A*, 547:A115.

Ivezic, Z., Tyson, J., Abel, B., et al., 2014. Lsst: from science drivers to reference design and anticipated data products. *arXiv preprint arXiv:0805.2366*.

J. Lotz, PI; M. Mountain, C.P., 2014. Hubble space telescope frontier fields.

Johnson, J., Karpathy, A., and Fei-Fei, L., 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Kalentev, O., Rai, A., Kemnitz, S., et al., 2011. Connected component labeling on a 2d grid using cuda. *Journal of Parallel and Distributed Computing*, 71(4):615.

Kartaltepe, J.S., Mozena, M., Kocevski, D., et al., 2015. Candels visual classifications: Scheme, data release, and first results. *The Astrophysical Journal Supplement Series*, 221(1):11.

Kennicutt Jr, R.C., 1998. Star formation in galaxies along the hubble sequence. *Annual Review of Astronomy and Astrophysics*, 36(1):189.

Kent, S.M., 1985. Ccd surface photometry of field galaxies. ii-bulge/disk decompositions. *The Astrophysical Journal Supplement Series*, 59:115.

Klusch, M. and Napiwotzki, R., 1993. HNS - a Hybrid Neural System and its Use for the Classification of Stars. *A&A*, 276:309.

Koekemoer, A.M., Faber, S., Ferguson, H.C., et al., 2011. Candels: The cosmic assembly near-infrared deep extragalactic legacy survey?the hubble space telescope observations, imaging data products, and mosaics. *The Astrophysical Journal Supplement Series*, 197(2):36.

Kormendy, J., 1977. Brightness distributions in compact and normal galaxies. iii-decomposition of observed profiles into spheroid and disk components. *The Astrophysical Journal*, 217:406.

Kormendy, J., Fisher, D.B., Cornell, M.E., et al., 2009. Structure and formation of elliptical and spheroidal galaxies. *The Astrophysical Journal Supplement Series*, 182(1):216.

Krizhevsky, A., Sutskever, I., and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kruk, S.J., Lintott, C.J., Bamford, S.P., et al., 2018. Galaxy zoo: secular evolution of barred galaxies from structural decomposition of multiband images. *Monthly Notices of the Royal Astronomical Society*, 473(4):4731.

Kruk, S.J., Lintott, C.J., Simmons, B.D., et al., 2017. Galaxy zoo: finding offset discs and bars in sdss galaxies. *Monthly Notices of the Royal Astronomical Society*, 469(3):3363.

Kulesza, T., Amershi, S., Caruana, R., et al., 2014. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3075–3084. ACM.

Kuminski, E. and Shamir, L., 2016. A computer-generated visual morphology catalog of 3,000,000 sdss galaxies. *The Astrophysical Journal Supplement Series*, 223(2):20.

Lahav, O., Naim, A., Buta, R.J., et al., 1995. Galaxies, Human Eyes, and Artificial Neural Networks. *Science*, 267:859.

Laureijs, R., Amiaux, J., Arduini, S., et al., 2011. Euclid definition study report. *arXiv preprint arXiv:1110.3193*.

Lazebnik, S., Schmid, C., and Ponce, J., 2005. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265.

LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *nature*, 521(7553):436.

LeCun, Y., Boser, B., Denker, J.S., et al., 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541.

Lee, D.D. and Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Leutenegger, S., Chli, M., and Siegwart, R.Y., 2011. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE.

Lindeberg, T., 1994. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225.

Lintott, C., Schawinski, K., Bamford, S., et al., 2010. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166.

Lintott, C., Schawinski, K., Bamford, S., et al., 2011. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166.

Lintott, C.J., Schawinski, K., Slosar, A., et al., 2008. 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.

Long, J., Shelhamer, E., and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

Lotz, J., Koekemoer, A., Coe, D., et al., 2017. The frontier fields: Survey design and initial results. *The Astrophysical Journal*, 837(1):97.

Lotz, J.M., Jonsson, P., Cox, T., et al., 2008. Galaxy merger morphologies and time-scales from simulations of equal-mass gas-rich disc mergers. *Monthly Notices of the Royal Astronomical Society*, 391(3):1137.

Lotz, J.M., Primack, J., and Madau, P., 2004. A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91.

Maaten, L.v.d. and Hinton, G., 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579.

Mendes, C., Gattass, M., and Lopes, H., 2013. Fgng: A fast multi-dimensional growing neural gas implementation. *Neurocomputing*.

Merlin, E., Amorín, R., Castellano, M., et al., 2016. The astrodeep frontier fields catalogues-i. multiwavelength photometry of abell-2744 and macs-j0416. *Astronomy & Astrophysics*, 590:A30.

Miller, A. and Coe, M., 1996. Star/galaxy classification using kohonen self-organizing maps. *Monthly Notices of the Royal Astronomical Society*, 279(1):293.

Miller, A.A., Bloom, J.S., Richards, J.W., et al., 2015. A Machine-learning Method to Infer Fundamental Stellar Parameters from Photometric Light Curves. *ApJ*, 798:122.

Mnih, V., Kavukcuoglu, K., Silver, D., et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Mortlock, A., 2013. Don't judge a galaxy by its cover.

Mortlock, A., Conselice, C.J., Hartley, W.G., et al., 2013. The redshift and mass dependence on the formation of the hubble sequence at z¿ 1 from candels/uds. *Monthly Notices of the Royal Astronomical Society*, 433(2):1185.

Muja, M. and Lowe, D.G., 2014. Scalable nearest neighbor algorithms for high dimensional data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36.

Naim, A., Ratnatunga, K.U., and Griffiths, R.E., 1997. Galaxy morphology without classification: Self-organizing maps. *The Astrophysical Journal Supplement Series*, 111(2):357.

Newell, A.J. and Griffin, L.D., 2011. Multiscale histogram of oriented gradient descriptors for robust character recognition. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1085–1089. IEEE.

Nielsen, M.L. and Odewahn, S.C., 1994. Automated Recognition of Galaxy Morphology Using Neural Networks. In *American Astronomical Society Meeting Abstracts*, volume 26 of *Bulletin of the American Astronomical Society*, page 107.09.

Noroozi, M. and Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer.

Odewahn, S.C., 1995. Automated Classification of Astronomical Images. *PASP*, 107:770.

Orlov, N., Shamir, L., Macura, T., et al., 2008. Wnd-charm: Multi-purpose image classification using compound image transforms. *Pattern recognition letters*, 29(11):1684.

Pearson, K., 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559.

Peng, C.Y., Ho, L.C., Impey, C.D., et al., 2002. Detailed Structural Decomposition of Galaxy Images. *AJ*, 124:266.

Petrosian, V., 1976. Surface brightness and evolution of galaxies. *The Astrophysical Journal*, 209:L1.

Raghavan, V., Bollmann, P., and Jung, G.S., 1989. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205.

Redmon, J., Divvala, S., Girshick, R., et al., 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788.

Ren, S., He, K., Girshick, R., et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Riccio, G., Cavuoti, S., Schisano, E., et al., 2016. Machine learning based data mining for milky way filamentary structures reconstruction. In *Advances in Neural Networks*, pages 27–36. Springer.

Roberts, M.S., 1963. The content of galaxies: Stars and gas. *Annual Review of Astronomy and Astrophysics*, 1(1):149.

Robitaille, T.P., Tollerud, E.J., Greenfield, P., et al., 2013. Astropy: A community python package for astronomy. *Astronomy & Astrophysics*, 558:A33.

Rublee, E., Rabaud, V., Konolige, K., et al., 2011. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE.

Samuel, A.L., 2000. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206.

Sandage, A., 2005. The classification of galaxies: Early history and ongoing developments. *Annu. Rev. Astron. Astrophys.*, 43:581.

Santos, J.M. and Embrechts, M., 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International Conference on Artificial Neural Networks*, pages 175–184. Springer.

Schawinski, K., Urry, C.M., Simmons, B.D., et al., 2014. The green valley is a red herring: Galaxy zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies. *Monthly Notices of the Royal Astronomical Society*, 440(1):889.

Schawinski, K., Zhang, C., Zhang, H., et al., 2017. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. *Monthly Notices of the Royal Astronomical Society: Letters*, 467(1):L110.

Schutter, A. and Shamir, L., 2015. Galaxy morphology?an unsupervised machine learning approach. *Astronomy and Computing*, 12:60.

Sculley, D., 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM.

Shallue, C.J. and Vanderburg, A., 2018. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. *The Astronomical Journal*, 155(2):94.

Shamir, L., 2011. Ganalyzer: A tool for automatic galaxy image analysis. *The Astrophysical Journal*, 736(2):141.

Shamir, L., 2012. Automatic detection of peculiar galaxies in large datasets of galaxy images. *Journal of Computational Science*, 3(3):181.

Shamir, L., Holincheck, A., and Wallin, J., 2013. Automatic quantitative morphological analysis of interacting galaxies. *Astronomy and Computing*, 2:67.

Shamir, L. and Wallin, J., 2014. Automatic detection and quantitative assessment of peculiar galaxy pairs in sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 443(4):3528.

Shu, F., 1982. *The physical universe: an introduction to astronomy*. University science books.

Silver, D., Huang, A., Maddison, C.J., et al., 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484.

Simard, L., 1998. Gim2d: an iraf package for the quantitative morphology analysis of distant galaxies. In *Astronomical Data Analysis Software and Systems VII*, volume 145, page 108.

Simard, L., Mendel, J.T., Patton, D.R., et al., 2011. A catalog of bulge+ disk decompositions and updated photometry for 1.12 million galaxies in the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 196(1):11.

Simmons, B.D., Lintott, C., Willett, K.W., et al., 2016a. Galaxy zoo: Quantitative visual morphological classifications for 48,000 galaxies from candels. *Monthly Notices of the Royal Astronomical Society*, page stw2587.

Simmons, B.D., Lintott, C., Willett, K.W., et al., 2016b. Galaxy zoo: Quantitative visual morphological classifications for 48000 galaxies from candels. *MNRAS*.

Skelton, R.E., Whitaker, K.E., Momcheva, I.G., et al., 2014. 3d-hst wfc3-selected photometric catalogs in the five candels/3d-hst fields: Photometry, photometric redshifts, and stellar masses. *The Astrophysical Journal Supplement Series*, 214(2):24.

Smith, S.W. et al., 1997. The scientist and engineer's guide to digital signal processing.

Sobel, I., 1990. An isotropic 3× 3 image gradient operator. *Machine vision for three-dimensional scenes*, pages 376–379.

Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11.

Sérsic, J.L., 1968. Atlas de galaxias australes. *Cordoba, Argentina: Observatorio Astronomico, 1968*.

Stoughton, C., Lupton, R.H., Bernardi, M., et al., 2002. Sloan digital sky survey: early data release. *The Astronomical Journal*, 123(1):485.

Sutskever, I., Vinyals, O., and Le, Q.V., 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sutton, R.S. and Barto, A.G., 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Taşdemir, K., Yalçin, B., and Yildirim, I., 2015. Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures. *Pattern Recognition*, 48(4):1465.

The Astropy Collaboration, Price-Whelan, A.M., Sipőcz, B.M., et al., 2018. The Astropy Project: Building an inclusive, open-science project and status of the v2.0 core package. *ArXiv e-prints*.

van den Bergh, S., 1960a. A preliminary liminosity classification for galaxies of type sb. *The Astrophysical Journal*, 131:558.

van den Bergh, S., 1960b. A preliminary luminosity clssification of late-type galaxies. *The Astrophysical Journal*, 131:215.

Van der Wel, A., Bell, E., Häussler, B., et al., 2012. Structural parameters of galaxies in candels. *The Astrophysical Journal Supplement Series*, 203(2):24.

Vikram, V., Wadadekar, Y., Kembhavi, A.K., et al., 2010. Pymorph: automated galaxy structural parameter estimation using python. *Monthly Notices of the Royal Astronomical Society*, 409(4):1379.

Vincent, P., Larochelle, H., Lajoie, I., et al., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371.

Vinh, N.X., Epps, J., and Bailey, J., 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM.

Vinh, N.X., Epps, J., and Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837.

Way, M.J. and Klose, C.D., 2012. Can Self-Organizing Maps Accurately Predict Photometric Redshifts? *PASP*, 124:274.

Willett, K.W., Galloway, M.A., Bamford, S.P., et al., 2016. Galaxy zoo: morphological classifications for 120 000 galaxies in hst legacy imaging. *Monthly Notices of the Royal Astronomical Society*, 464(4):4176.

Willett, K.W., Lintott, C.J., Bamford, S.P., et al., 2013. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, page stt1458.

Wright, D.E., Lintott, C.J., Smartt, S.J., et al., 2017. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society*, 472(2):1315.

Wright, D.E., Smartt, S.J., Smith, K.W., et al., 2015. Machine learning for transient discovery in pan-starrs1 difference imaging. *Monthly Notices of the Royal Astronomical Society*, 449(1):451.

Wu, K., Otoo, E., and Suzuki, K., 2009. Optimizing two-pass connected-component labeling algorithms. *Pattern Analysis and Applications*, 12(2):117.

Wynne, C., 1968. Ritchey-chretien telescopes and extended field systems. *The Astrophysical Journal*, 152:675.

York, D.G., Adelman, J., Anderson Jr, J.E., et al., 2000. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579.

You, Q., Jin, H., Wang, Z., et al., 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Zwicky, F., 1937. Nebulae as gravitational lenses. *Physical Review*, 51(4):290.