Astronomy and Computing 2 (2013) 67–73

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Automatic quantitative morphological analysis of interacting galaxies

Lior Shamir[a,*], Anthony Holincheck[b], John Wallin[c]

[a] Lawrence Technological University, Southfield, MI, USA
[b] George Mason University, St. Louis, MO, USA
[c] Middle Tennessee State University, Murfreesboro, TN, USA

ARTICLE INFO

Article history:
Received 8 February 2013
Accepted 10 September 2013

Keywords:
Galaxies: structure
Galaxies: evolution
Methods: analytical
Techniques: image processing

ABSTRACT

The large number of galaxies imaged by digital sky surveys reinforces the need for computational methods for analyzing galaxy morphology. While the morphology of most galaxies can be associated with a stage on the Hubble sequence, the morphology of galaxy mergers is far more complex due to the combination of two or more galaxies with different morphologies and the interaction between them. Here we propose a computational method based on unsupervised machine learning that can quantitatively analyze morphologies of galaxy mergers and associate galaxies by their morphology. The method works by first generating multiple synthetic galaxy models for each galaxy merger, and then extracting a large set of numerical image content descriptors for each galaxy model. These numbers are weighted using Fisher discriminant scores, and then the similarities between the galaxy mergers are deduced using a variation of Weighted Nearest Neighbor analysis such that the Fisher scores are used as weights. The similarities between the galaxy mergers are visualized using phylogenies to provide a graph that reflects the morphological similarities between the different galaxy mergers, and thus quantitatively profile the morphology of galaxy mergers.

1. Introduction

Galaxy mergers are linked to multiple forms of galactic activities such as star formation (Di Matteo et al., 2007; Bridge et al., 2007), quasars (Hopkins et al., 2005), active galactic nuclei (Springel et al., 2008), and galaxy morphology (Springel and Hernquist, 2005; Bower et al., 2006). While most single galaxies can be associated with a stage on the Hubble sequence, the morphology of a pair of interacting galaxies is more complex than the morphology of a single galaxy, making morphological analysis and classification of galaxy mergers a challenging task that requires compound catalogs and classification schemes (Arp, 1966; Struck, 1999; Schombert et al., 1990). The morphology of galaxies in these catalogs is determined by manual observation, and can therefore be ambiguous. Classes defined by Arp (1966) and by Vorontsov-Velyaminov (1959, 1977) differ in their characterization of interacting pairs. In some cases, even the class definitions can be ambiguous. For example the VV classes of pair of coalescents and pair in contact are difficult to distinguish. Even with simple galaxy morphology classifications, Galaxy Zoo has shown that some systems are intrinsically harder to classify into simple spiral and elliptical categories (Lintott et al., 2008).

Until now, automatic tools for galaxy morphological analysis have focused on single, non-interacting, galaxies. Current approaches include parametric model-driven methods such as GALFIT (Peng et al., 2002), GIM2D (Simard, 1998), and Ganalyzer (Shamir, 2011a,b), as well as machine learning methods (Ball et al., 2004, 2008; Shamir, 2009; Banerji et al., 2010; Huertas-Company et al., 2011). However, these methods are based on supervised machine learning, which automates a human guided classification of objects into one of several pre-defined distinct classes. GALFIT has been used with some success to model irregular and interacting galaxies (Peng et al., 2010) but only by fitting over 100 parameters in the best-fit models. Handcrafting these models for large numbers of interacting galaxies and those with irregular morphologies remains a daunting task. Other proposed methods include CAS (Conselice, 2003), and the Gini coefficient method (Abraham et al., 2003) that was also applied to galaxy images to deduce the statistics of the broad morphology of galaxy mergers (Lotz et al., 2008).

Unlike supervised machine learning, unsupervised machine learning is not based on existing knowledge and pre-defined training data, but aims at analyzing given data to automatically deduce its properties and structural descriptors. That is, in unsupervised learning the data are processed with no prior assumptions or human guidance to detect subsets of samples that are similar to each other, outliers, etc. In this paper we describe a method that can profile the morphology of interacting galaxies, and automatically deduce the similarities between galaxy mergers based on the galaxy

* Corresponding author. Tel.: +1 248 204 3512; fax: +1 248 204 3518.
E-mail address: lshamir@mtu.edu (L. Shamir).

2213-1337/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ascom.2013.09.002

images. Simulation remains an important tool for studying the morphology of interacting galaxies. The Zooniverse project sponsored the Merger Zoo to study dozens of interacting galaxies by having Citizen Scientist volunteers attempt to simulate specific pairs of galaxies. They were tasked with identifying and evaluating the results of galaxy simulations based on how well they matched the target images. The volunteers selected over 50,000 simulations during the course of the project. These simulated images form the population of training data used in this study.

2. Data

The data used in the experiment are 54 images of interacting galaxies taken by the Sloan Digital Sky Survey (Schneider et al., 2003). Additionally, the 50,000 simulated images of these galaxies produced by the Merger Zoo project are sampled for training data. The images were of size 512×512 pixels, downscaled to 256×256 to reduce the response time of the computation process.

3. Generating simulations of interacting galaxies

To effectively profile the morphology of a galaxy merger, there is a need for multiple images of each galaxy so that the pattern of morphological features can be deduced. For that purpose, multiple simulated galaxy mergers were generated for each of the galaxy images using a restricted three-body simulation code called SPAM. The restricted three-body approach uses a static galaxy potential and massless test particles to produce the tidal features seen in galaxy interactions. These models do not reproduce the gas dynamics or star formation associated with real interactions. However, they have been useful in modeling the orbital interactions and basic morphology of interacting systems. The computational speed of the code allows thousands of runs to be completed in the time required for a moderate resolution run using a tree code such as Gadget (Springel et al., 2000).

The initial conditions needed to simulate the gravitational forces in an encounter between two interacting galaxies cannot be fully determined from observational data alone. A multi-year Citizen Science project called Galaxy Zoo—Mergers was undertaken to enlist the help of volunteers in determining simulation initial conditions for the 54 SDSS galaxy mergers studied here. The SPAM code would run on sets of randomly selected initial conditions. The volunteers would indicate which simulations were a possible morphological match to the target galaxies.

Subsequent rounds of review and evaluation would assign a fitness score to each simulation based upon how well it matched the target image morphology. A perfect fit would be assigned a fitness of 1, and poor fits would have a fitness as low as 0. The sets of simulations used here all had a fitness of $\sim$0.21 or greater. Details of this project are contained in Holincheck (2013).

For this study, a set of simulated images were generated using 20,000 massless test particles. The particles were assigned to each galaxy in proportion to the specified masses. The final positions of the test particles were used to build up a grayscale image. For each particle, a Gaussian kernel was computed and the intensity of the pixel was set based upon the value of the kernel. Multiple particle activations were added together. The activation values for all of the pixels were then adjusted so that there were a total of 255 steps on a logarithmic scale between the lowest activation and the highest activation.

The use of simulated images to capture the morphological features of interacting systems is somewhat unusual. However, the Merger Zoo project has provided an opportunity to create a set of systems that share characteristics with the original galaxy. A more standard technique of scaling, rotating, and smoothing galaxy images would provide an alternative way of obtaining these data. However, the use of these models effectively allows the inclusion of the collisional history of these systems to be part of the classification process.

The output images are in the lossless TIFF format, and each image is rotated by a random number of degrees so that the machine learning will not be biased by the absolute positions of the two interacting galaxies, which are constant across all simulated galaxy mergers generated by SPAM for a certain image of interacting galaxies. Source code of the SPAM galaxy merger simulator is available for free download from the Astrophysics Source Code Library at http://coms.cs.mtsu.edu/jspam.

4. Unsupervised learning of galaxy morphology

Interacting galaxies feature complex morphology, and therefore comprehensive morphological analysis of galaxy mergers should be based on multiple numerical image content descriptors that reflect the image content. The feature set used in the analysis is the *WND-CHARM* scheme (Shamir et al., 2008a,b), which is based on extracting a very large number of image features, and was originally designed for automatic morphological analysis of cell and tissue images (Shamir et al., 2008b, 2009). The comprehensiveness of the feature set and its ability to measure very many different aspects of the visual content allows it to analyze complex morphology such as visual art (Shamir et al., 2010; Shamir and Tarakhovsky, 2012), and was also shown to be effective in the analysis of galaxy images (Shamir, 2009, 2012).

WND-CHARM first extracts from each image a vector of 1025 numerical image content descriptors that include high-contrast features (object statistics, edge statistics, Gabor filters), textures (Haralick, Tamura), statistical distribution of the pixel values (multi-scale histograms, first four moments), factors from polynomial decomposition of the image (Chebyshev statistics, Chebyshev–Fourier statistics, Zernike polynomials), Radon features and fractal features. These features are computed from the raw pixels, but also from image transforms as well as multi-order transforms. These transforms include the Fourier transform, Chebyshev transform, and Wavelet transform, as well as tandem combinations of these transforms. A detailed description of these image content descriptors and the image transforms is available in Shamir (2008), Shamir et al. (2008a,b, 2009, 2010) and Shamir (2012).

After the image features are computed, the simulated galaxy images of each target image are separated randomly into training and test sets such that 210 images are allocated for training and 20 for testing, and each of the 1025 features computed on the training set is assigned a Fisher discriminant score (Bishop, 2006). Since not all image content descriptors are expected to be informative, the features are ordered by their Fisher discriminant score, and 85% of the features with the lowest scores are rejected in order to filter non-informative image features.

The reason for computing the entire feature set before rejecting most features is that *WND-CHARM* is a data-driven algorithm, and therefore the informativeness of the features is determined statistically based on the data being processed. That is, *WND-CHARM* does not know which features are more informative before computing them, and therefore needs to compute all feature values for each image so that the most informative features can be selected. That approach of using a comprehensive set of numerical image content descriptors that reflect very many aspects of the visual content and then statistically selecting the most relevant features allows using the system without making any prior assumptions about the physical characteristics of the galaxies.

As described in Shamir et al. (2008a) and Orlov et al. (2008), the features are computed in groups, so that if a certain feature is needed the entire group needs to be computed, and there is no