Journal of Computational Science 3 (2012) 181–189

Contents lists available at SciVerse ScienceDirect

Journal of Computational Science

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

Automatic detection of peculiar galaxies in large datasets of galaxy images

Lior Shamir*

Lawrence Technological University, 21000 W Ten Mile Rd., Southfield, MI 48075, USA

ARTICLE INFO

Article history:
Received 28 September 2011
Received in revised form 6 March 2012
Accepted 9 March 2012
Available online 17 March 2012

Keywords:
Astroinformatics
Galaxies
Peculiar galaxies
Computational astrophysics
Knowledge discovery

ABSTRACT

We propose an image analysis unsupervised learning algorithm that can detect peculiar galaxies in datasets of galaxy images. The algorithm first computes a large set of calculated characteristics reflecting different aspects of the visual content, and then weighs them based on the σ of the values computed from the galaxy images. The weighted Euclidean distance of each galaxy image from the median is measured, and the peculiarity of each galaxy is determined based on that distance. Experimental results using irregular galaxy images show that the method can effectively detect peculiar galaxies. Code and data used in the experiments are freely available.

1. Introduction

Autonomous sky surveys are capable of generating very large datasets consisting of millions of galaxy images. However, in many cases datasets produced by high-throughput sky surveys such as SDSS (Sloan Digital Sky Survey) [1], Pan-STARRS (Panoramic Survey Telescope and Rapid Response System) [2], DES (Dark Energy Survey) [3], and LSST (Large Synoptic Survey Telescope) [4], can be too large to analyze manually. For instance, LSST is expected to collect ~30 TB of data each night, producing the world's largest public dataset.

One approach to mass analysis of galaxy images was taken by the Galaxy Zoo project [5], in which SDSS images were uploaded to a web site where thousands of amateur astronomers can manually classify the galaxy images. Due to the high number of volunteers that the venture attracted, almost a million galaxies were classified [6]. However, this subset of SDSS galaxies is merely a small fraction of SDSS DR8, which consists of ~200 M galaxies, and with ventures such as DES and LSST, which will see deeper space, these datasets are expected to grow much larger. Therefore, citizen science can provide just a limited solution to the problem of galaxy image analysis, and will not scale when applied to full existing and future large galaxy image datasets. This reinforces the need for automated methods that can effectively mine very large datasets of galaxy and turn these data into knowledge.

Automated methods for galaxy image analysis have focused on automatic classification of galaxy image datasets into several known morphological classes. Refs. [7–9], attempted to classify elliptical galaxies by analyzing their internal structure. Other attempts to perform automatic morphological classification of galaxies include the Gini coefficient method [10] and the CAS method [11]. Recently, machine learning algorithms were applied to automatic classification of galaxy images [12,13] and demonstrated considerable success in several experiments.

While machine learning algorithms have achieved classification accuracy of ~90% for the basic morphological classes of spiral, elliptical, and edge-on galaxies, these algorithms are based on supervised machine learning [12,13], or pre-defined galaxy morphology models [10,14], and require a step of training using pre-classified samples. Therefore, these supervised machine learning algorithms are limited to known galaxy morphological types, and are not designed to detect peculiar galaxies or morphological types on which the algorithms have not been trained.

It should be noted that while galaxy classification can be conceptualized as a classification problem, the classification of galaxy images is not a typical problem of classifying objects into one of several discrete classes, as the vast majority of galaxies in the universe do not fall into discrete categories but are placed on the continuous scale that reflects the process of galaxy morphology alteration, known as the Hubble sequence. However, while a vast majority of the galaxies observed through telescopes are indeed associated with a known class of galaxies or a stage on the Hubble sequence, some galaxies might be of forms that have not (or rarely) been observed before [15]. Naturally, these unique celestial objects are of high scientific interest.

* Tel.: +1 248 204 3512; fax: +1 248 204 3518.
E-mail address: lshamir@mtu.edu

1877-7503/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jocs.2012.03.004

An example of such object that was discovered in the SDSS galaxy image dataset is the object known as "Hanny's Voorwerp" [16,17], which was discovered by an amateur astronomer while she was classifying Galaxy Zoo images. The discovery of "Hanny's Voorwerp" demonstrates that it is reasonable to assume that astronomical objects of high scientific interest that have not been observed or characterized before might be hidden inside these large datasets. Since these objects are not known, it is required to develop and apply outlier detection methods that are effective for automatic detection of peculiar galaxies. Outliers are often not considered desirable in a scientific dataset [19,18]. However, in the case of galaxy datasets the peculiar galaxies that do not fall into any of the known galaxy morphological types are of high scientific interest. While autonomous sky surveys attract significant research resources, no such systems have been reported to successfully mine these galaxy image datasets for peculiar celestial objects. In Section 2 we describe the image analysis method, in Section 3 we discuss the image data that was used in the experiment, and in Section 4 the experimental results are presented using several different galaxy morphological types.

2. Image analysis method

The image analysis method used in this study is based on the CHARM feature set of the WND-CHARM scheme [20,21], which was originally developed for microscopy image analysis and classification of cells, but demonstrated efficacy also in supervised classification of galaxy images [12]. The CHARM feature set includes 2873 numerical image content descriptors that reflect the image textures, polynomial decomposition, statistical distribution of the pixel intensities, fractals, edges, and geometrical features, as described in [20–24]. The set of image content descriptors include the following algorithms:

1. Haralick texture features [25], which contribute 28 numerical texture descriptors based on the gray-level co-occurrence matrix as used by [26].
2. Tamura textures features, providing the contrast, directionality and coarseness of the texture [27]. Coarseness measures the scale of the texture, and is reflected by its 3-bin histogram and the sum of the 3-bin histogram. The contrast estimates the dynamic range of the pixel intensities, and directionality indicates whether the image is oriented toward a certain direction.
3. Gabor wavelets [28], computed using a Gaussian harmonic kernel and seven different frequencies of 1 through 7.
4. Multi-scale histograms of the pixel intensities using 3, 5, 7, and 9 bins, providing a total of 24 features normalized by the maximum number of counts.
5. First four moments of mean, standard deviation, skewness, and kurtosis of the pixels intensities computed on a set of 10 "stripes" in each of four different directions (0°, 45°, 90°, and 135°), and the values of the stripes of each direction are sampled into a 3-bin histogram. The four moments computed in four different directions using 3-bin histogram provides a total of 48 image content descriptors.
6. Zernike coefficients [29], used in the same fashion as described in [26], providing 72 image content descriptors.
7. Radon transforms computed in four different directions (0°, 45°, 90°, and 135°), and convolved into a 3-bin histogram. Three bins from each of the four directions produce a total of 12 numerical image content descriptors reflecting spatial information where pixels are correlated to a specific angle.
8. Object statistics computed by applying the Otsu binary mask [30] and then finding all 8-connected objects. Computed statistics include the Euler Number [31], which is the number of objects in the region minus the number of holes in those objects, as well as the minimum, maximum, mean, median, variance, and a 10-bin histogram of both the objects areas and the distances from objects centroids to the image centroid.
9. Chebyshev statistics computed using a 32-bin histogram of the coefficient of Chebyshev transform with the order of 20. The image features are the 32-bin histogram of the 400 coefficients [22,20].
10. Edge statistics computed by using the Prewitt gradient. The statistics includes the mean, median, variance, number of edge pixels in the image as a fraction of the total number of pixels, the direction homogeneity, and the 8-bin histogram of both the magnitude and the direction of the gradient.
11. Fractals features, as described by [32] and used for galaxy image analysis by [12].
12. The Gini coefficient as described by [10], and is informative for classifying galaxy images [10].

An interesting feature of the CHARM scheme is that the image features are extracted not just from the raw pixels, but also from several transforms of the image, which are the Fourier transform, Chebyshev transform, Wavelet transform (Symlet 5), and the edge transform, which is the magnitude component of the Prewitt gradient, binarized by the Otsu global threshold [30]. Image features are also extracted from tandem transforms, which include the Chebyshev transform followed by Fourier transform and Wavelet transform, Wavelet transform followed by Fourier transform, Fourier transform followed by Wavelet transform and Chebyshev transform, and edge transform followed by Fourier transform and Wavelet transform. The paths of the compound image transforms are described in [20,24], and a thorough analysis of the efficacy of using compound transforms is available in [33]. The Wndchrm source code is publicly available for free download at http://vfacstaff.ltu.edu/lshamir/downloads/ImageClassifier.

In the first stage of the algorithm, the values of all CHARM image features are normalized to the interval [0,1] based on the observed values computed from the images in the dataset, so that the differences between the values of different image features can be compared without introducing a numerical bias. For instance, if the values of one image feature are in the range of [0,1000] while the values of another feature are in the range of [0,10], a numerical difference of 8 between the values of the first feature extracted from two different images can be considered small, while the same numerical difference can be much more substantial for the second feature, in which it is almost the entire range. Fig. 1 shows the histograms of three of the 2873 image features used in [12], which are bin 31 of the Zernike features, bin 4 of the Chebyshev statistics, and bin 7 of the Haralick texture features computed from the Chebyshev transform. As the figure shows, although the numerical descriptors are considered informative for galaxy classification based on morphology compared to the other features [12], the differences between the morphological types reflected by the histograms are not strong. Therefore, a single numerical image content descriptor in the feature set cannot be effectively used for automatic analysis of galaxy morphology, and a combination of features should be used in concert to provide useful automatic analysis of the morphology.

The vast majority of galaxies in the universe observed from Earth fall into one of a few morphological categories, which are spiral galaxies, elliptical/S0 galaxies, and edge-on galaxies of which the visual morphology cannot be determined from Earth. Galaxies that do not fall into these categories are considered "peculiar", and might be of high scientific interest. The fact that galaxy morphology in the universe is relatively homogenous makes it easier to detect peculiar galaxies by using the distribution of the image content descriptor values, and the histograms of many informative image