Algorithmic and machine learning approaches to automatic identification of peculiar galaxies in large astronomical databases

Lior Shamir[1]

[1], *Kansas State University, Manhattan, KS, United States;* `lshamir@mtu.edu`

Abstract. Digital sky surveys can image many millions of extra-galactic objects. While the majority of these objects are galaxies of known types, a small portion of them have rarely or never seen before. These objects possess critical information about the past, present, or future Universe. Since they are hidden inside very large databases, finding them is impractical by manual analysis, and therefore automation is required. Different approaches for automatic identification include model-driven and data-driven approaches, and can be further separated into supervised and unsupervised machine learning. Due to the extreme size of these databases, even a mild rate of false positives will make the output unmanageable. Therefore, a practical solution to the detection of peculiar objects must be able to control the false-positive rate, while also handling artifacts or saturated images, which are common in digital sky surveys. The algorithms, approaches, and examples of detected objects are described, with application to data from DES, SDSS, and HST.

1. Introduction

Autonomous digital sky surveys have enabled the acquisition of large astronomical databases, containing millions, and sometimes billions of celestial objects. Most of these objects are of known types. But in databases containing many millions of objects, it can be assumed that many of these objects are rare or novel to science. Hidden among a very large number of "regular" objects, the only practical way of finding these objects is by applying automation.

In the pre-information era, catalogs of peculiar galaxies were collected manually. An example of such catalog is the Atlas of Peculiar Galaxies (Arp and Madore 1975). That catalog enabled numerous studies, but it required over a decade of work to compile. Digital sky surveys collect a large number of galaxy images, but while they make the data easily accessible, identifying these objects among millions of "regular" objects is a challenging task.

Some previous work was based on manual browsing through the image data (Kaviraj 2010). That also included catalogs of known but rare objects such as collisional ring galaxies (Madore et al. 2009), or other ring galaxies (Finkelman et al. 2012; Buta 2017). Algorithms based on machine learning were also proposed and applied to detect galaxy mergers (Margalef-Bentabol et al. 2020), peculiar galaxy pairs (Shamir and Wallin 2014), ring galaxies (Timmis and Shamir 2017; Shamir 2020), and gravitational lenses (Jacobs et al. 2019). This paper discusses approaches and methods for automatic identification of peculiar galaxies in digital sky surveys.

2. Model-driven approach for detection of known morphological types

When rare galaxies have known and defined shape, it is often possible to develop simple algorithms to identify them. That can be done by either training a supervised machine learning algorithm, or by model-driven algorithms that are based on computer vision elements. An example is the automatic identification of ring galaxies. Ring galaxies are well-known objects, yet relatively uncommon compared to other types of galaxies. Because ring galaxies have known and defined shape, it is possible to detect them by either training a machine learning system, or by a model-driven analysis that searches for rings in the galaxy shape. A simple yet effective identification in real-world data is demonstrated in (Timmis and Shamir 2017; Shamir 2020), providing catalogs of ring galaxies from Pan-STARRS (Timmis and Shamir 2017) and SDSS (Shamir 2020) data.

The algorithm works by first applying dynamic thresholding to separate the background pixels from the foreground of the galaxy. For each threshold, the algorithm scans all background pixels and applies a flood-fill algorithm to attempt to find a surface that stretches from the background pixel to the edge of the image. If the surface does not reach any of the image edges, the size of the surface (in pixels) is compared to the number of foreground pixels. The presence of an area that is sufficiently large and contained inside the galaxy foreground indicate that the galaxy has ring-like features.

Although the algorithm is simple, and it is not complete in the sense that it does not recognize all ring galaxies, it is fast and can identify ring galaxies in very large databases of real-world images. For instance, Figure 1 shows examples of ring galaxies detected in Pan-STARRS. That catalogs of ring galaxies (Timmis and Shamir 2017; Shamir 2020) require substantial labor to compile manually.

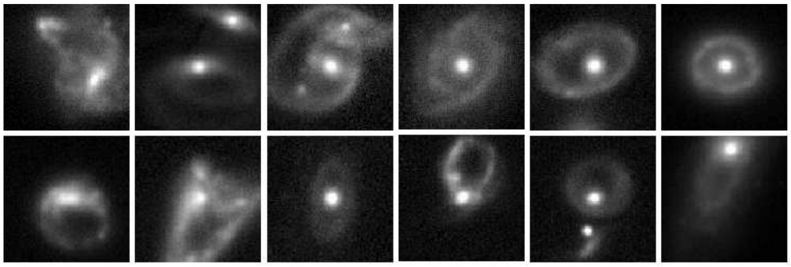

Figure 1. Example ring galaxies identified in Pan-STARRS (Timmis and Shamir 2017).

3. Unsupervised machine learning for the identification of unknown morphological types

Very large databases can contain many objects of types that are not yet known to science. Because they are not known, it is not possible to develop model-driven algorithms to detect them. It is also not possible to train a supervised machine learning system, as images of such objects do not yet exist. One of the approaches to detect such objects is by applying unsupervised machine learning.

When using deep convolutional neural networks, autoencoders is a common approach for novelty detection in image data. An example of the application of autoencoders to identify peculiar galaxies in image data is (Margapuri et al. 2020). The autoencoder reconstruction loss is used as an indication of the difference between the tested image and the other images in the database, and a higher reconstruction loss might indicate that the image is an outlier galaxy.

Another approach to the detection of outlier galaxies is to first compute numerical image content descriptors from each galaxy image, and then apply outlier detection algorithms to these descriptors (Shamir 2012). An important advantage of this "shallow learning" approach is its ability to control of the false-positive rate. Given the large size of databases collected by modern digital sky surveys, even a small rate of false-positives can result in a very large number of objects that need to be observed manually. Figures 2, 3, and 4 show examples of outliers detected automatically in SDSS, HST, and DES, respectively. A complete list of the objects can be found in (Shamir and Wallin 2014; Shamir 2021).

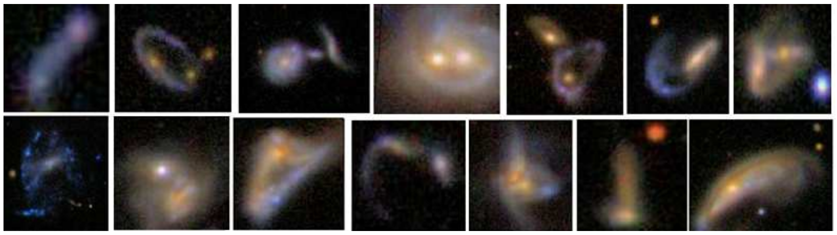

Figure 2. Outlier extended objects detected in SDSS (Shamir and Wallin 2014).

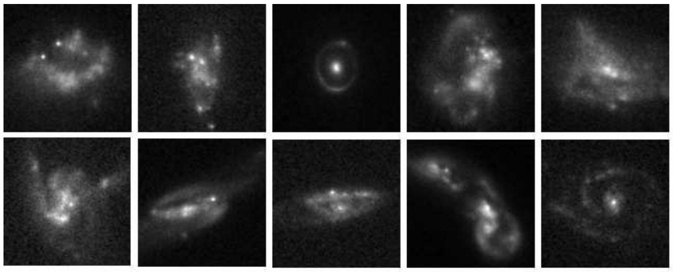

Figure 3. Examples of outlier objects detected automatically in HST (Shamir 2021).

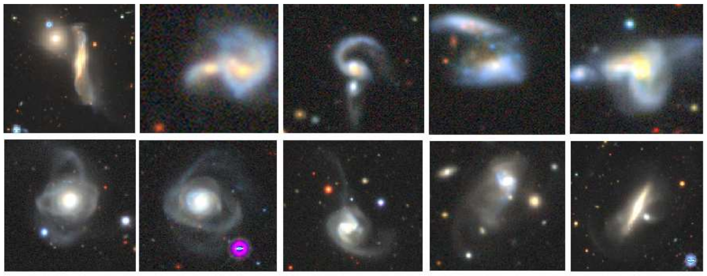

Figure 4. Examples of outlier objects detected in the Dark Energy Survey.

The algorithm works by first extracting a comprehensive set of numerical image content descriptors from each galaxy image. These descriptors include textures, edges, fractals, polynomial decomposition, statistical distribution of the pixel intensities, and numerous transforms of the images as described in detail in (Shamir et al. 2008, 2010).

After the numerical image content descriptors are computed, they are ranked by their Boltzmann entropy to select the features that are the most informative. Then, the distances between all pairs of images are computed by using the Earth Mover's Distance (Rubner et al. 2000), where the entropy values computed for the features are used as weights (Shamir 2021). The Earth Mover's Distance (EMD) between the pairs of images reflect the similarity between them, such that lower distance means that the two images are more similar to each other.

Once the distances are computed, an outlier galaxy can be identified by ranking the EMD distances between that galaxy and all other galaxies in the dataset. Instead of using the shortest distance between the galaxy and any of the other galaxies, the method works by using the N*th* shortest distance, such as $N \geq 1$. The N*th* shortest EMD is used because in very large datasets of galaxy images rare galaxies of the same type can appear more than once. That can lead to several galaxy images that are similar to each other, but are different from all other galaxy images. A short distance between any two galaxies might therefore reflect two or more outlier galaxies that are similar to each other but could be different from all other galaxies. In that case, these galaxies will not be detected as outliers. Additionally, using just the one shortest distance might lead to a high number of false-positives, and mainly saturated images and artefacts.

For these reasons, using the N*th* distance can allow to avoid some of these saturated images that are common in the digital sky survey data. Controlling the distances and the number of neighbors used allows to adjust the trade-off between false positives and the sensitivity of the algorithm. As mentioned above, even a small rate of false positives can make the algorithm practically unusable. The algorithm is not complete, and has a certain rate of false positives. Full analysis can be found in (Shamir 2021). But the algorithm reduces the data by three orders of magnitudes, making the detection of novelty objects practical (Shamir and Wallin 2014; Shamir 2021).

References

Arp, H. C. and Madore, B. F. (1975). A catalogue of southern peculiar galaxies from the uk schmidt survey: Preliminary reductions of 36 fields. *The Observatory*, 95:212–214.

Buta, R. J. (2017). Galactic rings revisited–i. cvrhs classifications of 3962 ringed galaxies from the galaxy zoo 2 database. *MNRAS*, 471(4):4027–4046.

Finkelman, I., Funes SJ, J. G., and Brosch, N. (2012). Polar ring galaxies in the galaxy zoo. *MNRAS*, 422(3):2386–2398.

Jacobs, C. et al. (2019). Finding high-redshift strong lenses in des using convolutional neural networks. *MNRAS*, 484(4):5330–5349.

Kaviraj, S. (2010). Peculiar early-type galaxies in the sloan digital sky survey stripe82. *MNRAS*, 406(1):382–394.

Madore, B. F., Nelson, E., and Petrillo, K. (2009). Atlas and catalog of collisional ring galaxies. *ApJS*, 181(2):572.

Margalef-Bentabol et al. (2020). Detecting outliers in astronomical images with deepgenerative networks. *arXiv:2003.08263*.

Margapuri, V. S. K., Shamir, L., and Thapa, B. (2020). Detection of unknown galaxy types in large databases of galaxy images. In *29th International Conference on Software Engineering and Data Engineering*. ISCA.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121.

Shamir, L. (2012). Automatic detection of peculiar galaxies in large datasets of galaxy images. *Journal of Computational Science*, 3(3):181–189.

Shamir, L. (2020). Automatic detection of full ring galaxy candidates in sdss. *MNRAS*, 491(3):3767–3777.

Shamir, L. (2021). Automatic identification of outliers in hubble space telescope galaxy images. *MNRAS*, 501(4):5229–5238.

Shamir, L. et al. (2008). Wndchrm–an open source utility for biological image analysis. *Source Code for Biology and Medicine*, 3(1):13.

Shamir, L. et al. (2010). Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7(2):1–17.

Shamir, L. and Wallin, J. (2014). Automatic detection and quantitative assessment of peculiar galaxy pairs in sloan digital sky survey. *MNRAS*, 443(4):3528–3537.

Timmis, I. and Shamir, L. (2017). A catalog of automatically detected ring galaxy candidates in panstarss. *ApJS*, 231(1):2.