Galaxy morphology - an unsupervised machine learning approach

Andrew Schutter, Lior Shamir*

Lawrence Technological University, Southfield, Michigan 48075, USA

Abstract

Structural properties posses valuable information about the formation and evolution of galaxies, and are important for understanding the past, present, and future universe. Here we use unsupervised machine learning methodology to analyze a network of similarities between galaxy morphological types, and automatically deduce a morphological sequence of galaxies. Application of the method to the EFIGI catalog show that the morphological scheme produced by the algorithm is largely in agreement with the De Vaucouleurs system, demonstrating the ability of computer vision and machine learning methods to automatically profile galaxy morphological sequences. The unsupervised analysis method is based on comprehensive computer vision techniques that compute the visual similarities between the different morphological types. Rather than relying on human cognition, the proposed system deduces the similarities between sets of galaxy images in an automatic manner, and is therefore not limited by the number of galaxies being analyzed. The source code of the method is publicly available, and the protocol of the experiment is included in the paper so that the experiment can be replicated, and the method can be used to analyze user-defined datasets of galaxy images.

Keywords: galaxies: structure – galaxies: evolution – methods: analytical – techniques: image processing

1 Introduction

In the past few years, advancements in computational tools and algorithms have started to allow automatic analysis of galaxy morphology. Approaches to automatic galaxy classification include model-driven methods such as GALFIT (Peng et al., 2002), GIM2D (Simard, 1999; Simard et al., 2011), CAS (Conselice, 2003), Gini(Abraham et al., 2003), Ganalyzer (Shamir, 2011), and SpArcFiRe (Davis and Hayes, 2014). Data-driven methods include binary classifiers that can differentiate between broad galaxy morphological types of elliptical and spiral galaxies (Shamir, 2009; Meneses Cuadros et al., 2009; Banerji et al., 2010), but also classifiers that can differentiate between four basic objects (Abd Elfattah et al., 2013), classification between four basic Hubble morphological types of E, S0, Sab, and Scd (Huertas-Company et al., 2010), and comprehensive analysis of galaxy images that include specific morphological features (Baillard et al., 2006; Kuminski et al., 2014; Dieleman et al., 2015). Classification of galaxies can also be performed using spectra in supervised (Ball et al., 2004) and unsupervised (Almeida et al., 2010) manner.

While supervised machine learning have demonstrated reasonable efficacy in automatic classification of galaxies by their morphological types (Shamir, 2009; Meneses Cuadros et al., 2009; Banerji et al., 2010; Huertas-Company et al., 2010), discrete classifiers do not effectively conceptualize the continuous nature of galaxy morphology, and therefore galaxy morphological schemes are still defined by manual observation. One of the earlier and most widely used schemes is the Hubble sequence (Hubble, 1936; Sandage, 1961), which is a commonly used morphology classification scheme that covers the morphology of most known galaxies. Hubble's initial work proposed a morphology classification system based on attributes of observed nebulae, originally consisting of three main morphological types, commonly known as elliptical (E), normal spirals (S) and barred spirals (SB) (Hubble, 1936). Humason et al. (1956) revisited the Hubble Sequence, introducing lenticular galaxies (S0), creating what is most commonly known as the Hubble “tuning-fork” diagram (De Vaucouleurs, 1959). It should be noted that although irregular (I) galaxies were recognized by Hubble, they were not included in Hubble’s classification scheme since at the time they could not be distinctively classified (De Vaucouleurs, 1959).

Since the Hubble morphological scheme was introduced, several modifications and enhancements have been proposed. Morgan and Mayall (1957), proposed a galaxy classification scheme based on the spectra, showing the correlation between the spectra and the spiral structure and spectral concentration (Morgan, 1958), and identified the cD phenomena (Morgan and Lesh, 1965). van den Bergh (1960) proposed a classification system of late-type galaxies based on luminosity, driven by the correlation between absolute luminosity and the shape of the spiral arms. That work was followed by a galaxy classification scheme of spiral and S0 galaxies, and distinguished “early” and “late” type systems by their disk-to-bulge ratio (Van Den Bergh, 1976).

Sandage (1961) showed that S0$_1$ to S0$_3$ galaxies do not feature an increase in flattening of the galaxies, and that normal spiral galaxies and S0 galaxies form two parallel sequences (Sandage et al., 1970; Van Den Bergh, 1976). Kormendy and Bender (1996) expanded the Hubble classification scheme with a more detailed morphological analysis of elliptical galaxies (Kormendy and Djorgovski, 1989). Kormendy and Bender (1996) proposed some modifications to the Hubble sequence, including the two-component S0 galaxies, and the addition of the Magellanic irregulars.

One of the notable refinements and extensions to the Hubble sequence was proposed by De Vaucouleurs (1959), proposing a three dimensional system. This classification included the four main broad morphological classes of elliptical, lenticular, spiral, and irregular galaxies along a linear main axis from galaxy types E to I$_m$, including Hubble’s initial a, b, c representation for ”early” to ”late”. Sandage (1961) refinement included d for ”very late” and the division of S0 galaxies into SO$^-$, SO^0, SO$^+$, as well as the inclusion of ”m” for magellanic galaxies: E, E$^+$, S0$^-$, S0^0, S0$^+$, S$_a$, S$_b$, S$_c$, S$_d$, S$_m$, or I$_m$ (De Vaucouleurs, 1994). The classification was also extended to include intermediate stages between the initial a, b, c, d, and m stages such as ab, bc, cd, and dm. This scheme introduced a notation based on *family*, *variety*, and *stage*, with *family* representing the absence of bars in a spiral galaxy (A), the presence of bars (B), or a transition of the two (AB). *Variety* represents the presence of a ring shape (r), spiral shape (s), or transition of the two (rs) within spiral galaxies, and *stage* represents the galaxy position along the main axis. Another feature of this classification scheme was assigning each *stage* along the main axis a numerical integer value between -6 and 11. E galaxies being represented by the values -6 to -4, lenticular -3 to -1, spirals 0 to 9, and irregulars 10 to 11 for a more quantitative approach to the classification. Furthering the quantitative approach to galaxy classification, De Vaucouleurs (1994) also introduced measurable parameters showing either a consistent mean increase or decrease along the current classification sequence. Characteristics include bulge-to-disk ratios, integrated luminosity in the B-band, the ratio of aperture diameters, total or effective magnitudes, mean surface brightness, and hydrogen index (De Vaucouleurs, 1994).

While proposing a quantitative scheme, the association of a galaxy to a morphological type is subjective, and the annotations of two or more astronomers are not necessarily identical in all cases (Naim et al., 1995; de Lapparent et al., 2011). It has been therefore proposed that galaxy morphology classification schemes will involve computational methods (De Vaucouleurs, 1994). Here we perform automatic unsupervised analysis of galaxy images of different morphological types to produce a computer-generated galaxy morphology sequence. The scheme is based on quantitative computer analysis of thousands of annotated galaxy images, producing a network of similarities between the morphological types that is independent of the human perception and the way humans quantify the similarities between these types.

2 Data

The data used in the study are taken from the EFIGI catalog (Baillard et al., 2011; de Lapparent et al., 2011), which was compiled for the purpose of developing and testing computational methods related to galaxy morphology. The catalog contains image data as well as morphological annotation data of 4458 galaxies taken from PGC (Principal Galaiesy Catalogue), also included in SDSS (Sloan Digital Sky Survey) Data Release 4. Among other morphological features, each galaxy was assigned with its morphological type determined by 10 astronomers (Baillard et al., 2011) based on the updated RC3-based Hubble types (De Vaucouleurs et al., 1992). Other morphological features include the bulge, spiral arms, as well as other features such as texture, appearance in the sky, and environment.

EFIGI contains images of each galaxy in the u, g, i, r, and z bands (Baillard et al., 2011). To produce color images, the i, r, and g bands were combined to provide a composite RGB image, such that gamma correction of 1.3 was applied to the luminosity, and color saturation was increased by a factor of 2. The color images were converted to the PNG (Portable Network Graphics) format using the STIFF software (Baillard et al., 2011).

The EFIGI color images were converted to 255×255 color 24-bit TIFF (Tagged Image File Format) images using *ImageMagick*, and were separated into folders such that each folders contained galaxies of the same type as annotated by EFIGI. Images of the same galaxies in the u, g, r, i, and z bands were converted to monochrome TIFF, and were used without color information.

The galaxy types are based on the numerical scheme (De Vaucouleurs, 1959) taken from the EFIGI catalog (Baillard et al., 2011). Each galaxy type had at least 142 samples, except for cE (-6), cD (-4), and dE (11), which only had 18, 44, and 69 samples, respectively. For their small size, these classes were not used in the experiment.

3 Image analysis method

The image analysis method used in the experiment is Wndchrm (Shamir et al., 2008a; Shamir, 2008; Shamir et al., 2009b, 2010a, 2013a), that has a feature set of 4027 numerical image content descriptors, or 2885 numerical descriptors when color information is not used. The numerical image content descriptors are the following:

Texture features:
1. **Haralick textures**: Energy and entropy computed on the co-occurrence matrix of the image (Haralick et al., 1973), measured using 28 image descriptor values as described in Shamir et al. (2008a).
2. **Tamura textures**: *Contrast*, *directionality* and *coarseness* of the Tamura textures (Tamura et al., 1978). The coarseness descriptors are its sum and its 3-bin histogram, providing a total of six numerical content descriptors.
3. **Gabor Filters**: Gabor filters (Gabor, 1946) using seven frequencies (1 through 7) and Gaussian harmonic function (Grigorescu et al., 2002).

Polynomial decomposition:
1. **Radon transform features** (Lim, 1990): Four series computed for angles 0, 45, 90, 135 degrees, and then convolved into a 3-bin histogram, providing a total of 12 numerical content descriptors.
2. **Chebyshev Statistics** (Gradsteyn and Ryzhik, 1994): A 32-bin histogram of a 400-bin vector produced by the Chebyshev transform of the with order of N=20.
3. **Zernike features**: Absolute values of the 72 coefficients of the Zernike polynomial approximation (Teague, 1980).
4. **Chebyshev-Fourier features**: A 32-bin histogram of the polynomial coefficients of a Chebyshev–Fourier transform (Orlov et al., 2008) with maximum polynomial order of N=23.

High-contrast features:
1. **Fractal features**, as described in (Wu et al., 1992).
2. **Edge features**: Mean, median, variance, and 8-bin histogram of the magnitude and direction computed on the Prewitt gradient of the image, as well as edge direction homogeneity.
3. **High-contrast object statistics**: Minimum, maximum, mean, median, variance, Euler number, and 10-bin histogram of the objects areas computed on the 8-connected objects found in the Otsu binary transform of the image.

Pixel statistics:
1. **Multi-scale Histograms**: Four histograms with 3, 5, 7, and 9 bins computed on the pixel intensities (Hadjidemetriou et al., 2001).
2. **First 4 Moments**: Mean, standard deviation, skewness, and kurtosis computed on image ”stripes” in four different directions (0, 45, 90, and 135 degrees).

These features are extracted not just from the raw values, but also from the two-dimensional transforms and combinations of multi-order transforms. The transforms are Fourier transform, Chebyshev transform, Wavelet (symlet 5, level 1) transform, color transform (Shamir, 2006), and edge magnitude transform. A detailed description and performance analysis of the image features extracted from image transforms and multi-order transforms can be found in (Shamir et al., 2008a; Shamir, 2008; Shamir et al., 2009b, 2010a, 2013a).

The comprehensive nature of the numerical image content descriptors allows analyzing complex morphology such as radiology (Shamir et al., 2009b,a), microscopy (Shamir et al., 2008b; Manning and Shamir, 2014), and visual art (Shamir et al., 2010a; Shamir, 2012b). In particular, the Wndchrm feature set has been proved to be informative for analysis of galaxy morphology, and was found useful for tasks such as galaxy classification (Shamir, 2009; Kuminski et al., 2014) and automatic detection of peculiar galaxies (Shamir, 2012a; Shamir and Wallin, 2014; Shamir et al., 2014a). A complete and detailed description of the set of numerical content descriptors and comprehensive performance analysis is available in (Shamir et al., 2008a; Orlov et al., 2008; Shamir, 2008; Shamir et al., 2010a, 2009b,a), and the source code is also publicly available through the Astrophysics Source Code Library (Shamir et al., 2013b).

As mentioned in Section 1, the purpose of the method is not to automatically classify galaxies by their morphology, but to quantitatively deduce a network of similarities between the different morphological types using merely the galaxy images, and without using metadata or existing knowledge that is not in the image content. The unsupervised analysis (Shamir et al., 2010a; Shamir and Tarakhovsky, 2012; Shamir et al., 2013a) works by first allocating 140 galaxy images from each galaxy type as annotated by EFIGI to the training set, and assigning each numerical image content descriptor with its Fisher discriminant score (Bishop et al., 2006) computed using the training samples. After the content descriptors were ranked based on their Fisher discriminant scores, the 85% of the least informative features, with the lowest Fisher scores, are rejected (Shamir, 2009; Shamir et al., 2009b,c).

The similarity between each pair of galaxy images is then computed using the Weighted Nearest Distance (WND) algorithm (Shamir et al., 2008a, 2010a). The mean similarity between all test galaxies of type t_1 and all training galaxies of type t_2 determines the similarity between these two galaxy morphological types. The similarities between all pairs of galaxy types produce a similarity matrix, normalized such that the similarity between a certain type to itself is set to 1 (Shamir et al., 2008a, 2010a; Shamir and Tarakhovsky, 2012; Shamir et al., 2013a). The similarity matrix is computed 20 times such that in each run different images are randomly allocated to training and test sets, and the final similarity matrix is generated by averaging the 20 similarity matrices.

The similarity matrix is visualized by PHYLIP (Felsenstein, 1993; Kuhner and Felsenstein, 1994), which was originally developed for visualizing similarities between organisms by their genotypes, but in this experiment used to visualize the similarities between galaxy types. It is used with randomized input order of sequences where 97 is the seed, 10 jumbles, and the Equal-Daylight arc optimization. When pairs of nodes are added, new nodes are created to provide the optimal tree that represents the similarity matrix. PHYLIP first creates the tree in the form of a text file that follows the Newick format, and then visualizes it by using the DRAWTREE program. The edges between the nodes reflect the degree of similarities between them, such that a shorter path between two nodes reflects a higher similarity between the images of these two classes. DRAWTREE automatically sets the angles such that the tree is convenient and easy to read. In the phylogeny created by PHYLIP each pair of nodes has just one possible path between them, and the length of the path includes all segments on that path, including edges between nodes added by PHYLIP during the tree optimization process.

The method used to compute and visualize the similarities between the galaxy types is described in details in (Shamir et al., 2010a; Shamir and Tarakhovsky, 2012; Shamir et al., 2014b), and was used for unsupervised analysis of simulated images of galaxy mergers (Shamir et al., 2013a). It also demonstrated its ability to profile continuous biomedical processes in which the clinical stages are reflected by image morphologies that change on a continuous scale (Shamir et al., 2010b). Detailed instructions including specific command lines used to produce the results are described in A.

4 Results

The application of the similarity estimation method described in Section 3 to the EFIGI color image data described in Section 2 produced the phylogeny displayed by Figure 1.

As the figure shows, the network of similarities between the galaxy morphological types computed by the algorithm is in agreement with the ordered sequence proposed by De Vaucouleurs (1959). The algorithm produced a graph starting with the ellipticals (-5), followed by the lenticualr galaxies (-3 through -1). Then, continuing sequentially are the spiral galaxies from (1 through 9) followed by the irregulars (10) in an order with perfect agreement with (De Vaucouleurs, 1959).

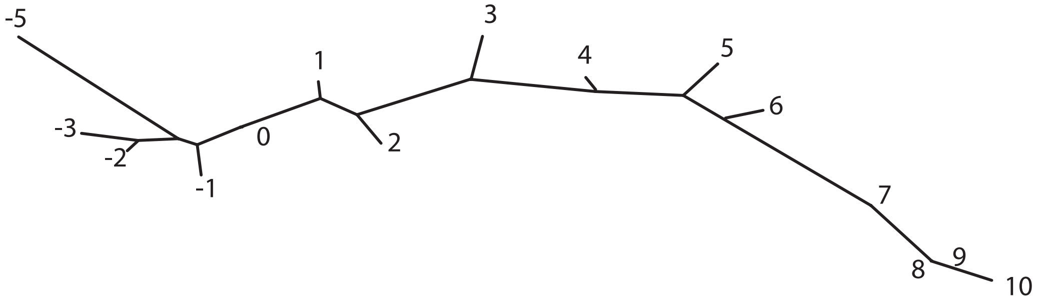

Figure 1: The network of similarities between the galaxy morphological types as deduced automatically by the algorithm

As mentioned in Section 2, the cD (-4), cE (-6), and dE types (11) were not included in the analysis due to the insufficient amount of sample images of these types in EFIGI. The probability that 15 elements are ordered in an ascending or descending order by mere chance is $\frac{2}{15!} =\sim 1.53 \cdot 10^{-12}$.

In another experiment we tested the method using the color images converted to gray-scale, and normalized for intensity such that all images had mean pixel value of 100, and standard deviation of 25 (Shamir et al., 2008a). The normalization ensured that the order will be determined by the shape, with no impact of color or brightness. The resulting graph produced by the algorithm is displayed in Figure 2.

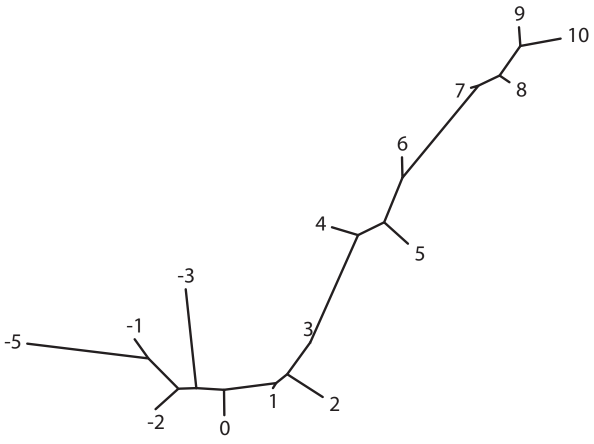

Figure 2: The network of similarities between the galaxy types using normalized gray-scale images

As the figure shows, the analysis of the normalized gray-scale images provided results similar to the graph produced using the color images, showing that the order was not necessarily driven by pixel intensity or by the color. The random chance probability that 12 elements out of 15 are ordered in ascending or descending order is $2 \cdot \binom{15}{12} \frac{1}{12!} =\sim 1.9 \cdot 10^{-6}$.

It is also noticeable that the S0 galaxy types S0^{-} (-3), S0^{0} (-2), and S0^{+} (-1) do not follow the numerical order proposed by De Vaucouleurs (1959). That analysis of the computer is in agreement with the observation that S0^{-}, S0^{0}, and S0^{+} galaxies do not feature an increase in the flattening of the galaxies (Sandage, 1961).

Figure 3 shows the Fisher discriminant scores of the groups of numerical image content descriptors, reflecting the measured informativeness of the descriptors and consequently their impact on the analysis. The descriptors are extracted from the image transforms and multi-order transforms.

As the figure shows, the identification of the Hubble stage depends on numerous image content descriptors working in concert. The fractal features were the most informative descriptors, indicating that the fractality of the galaxy is different across different galaxy morphological types. This agrees with the observation that fractality can be used as a galaxy classification signature (Lekshmi et al., 2003), and can assist in differentiating between elliptical and spiral galaxies (Shamir, 2009). For instance, an elliptical galaxy has low fractality in the absence of complex shape, but the fractality of a galaxy should become more dominant when the galaxy has more arms and split arms.

The graph shows that many other numerical image content descriptors such as Haralick textures (Haralick et al., 1973) have an impact on the analy-

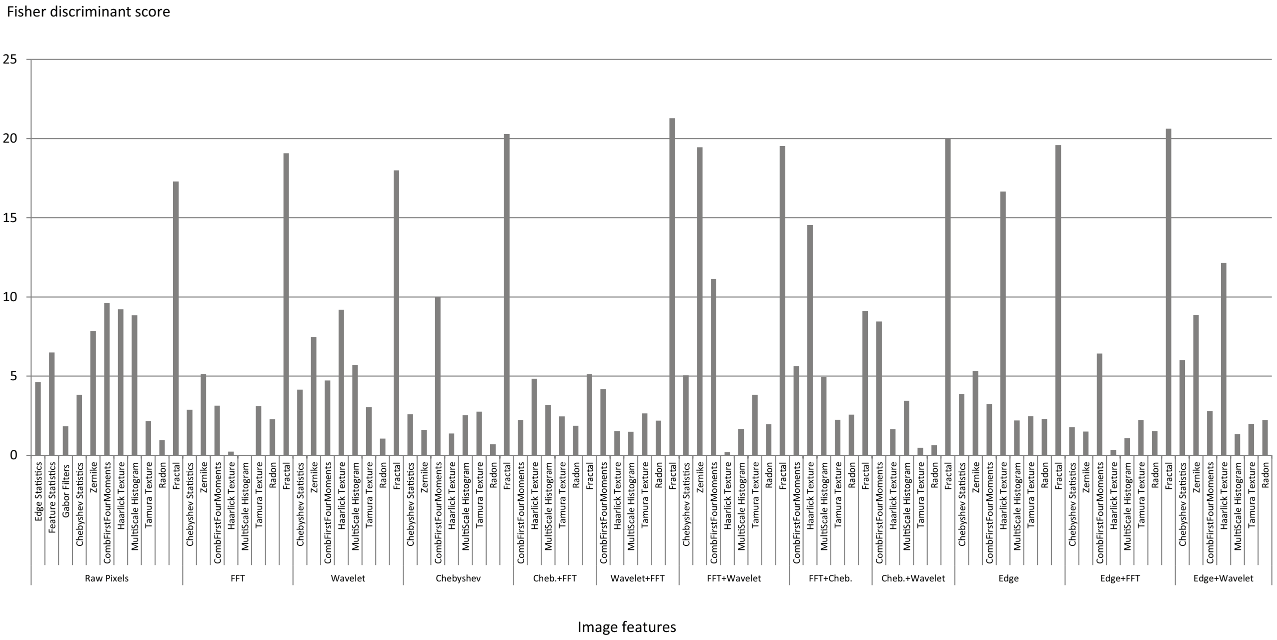

Figure 3: Fisher discriminant scores of the different groups of numerical content descriptors, extracted from the different image transforms

sis, and work in concert to quantify the similarities between the different galaxy morphological types. Texture features has been shown to be informative in separating between galaxies based on their morphological types (Au, 2006; Shamir, 2009; Banerji et al., 2010; Pedersen et al., 2013). For instance, texture homogeneity/entropy may change as the galaxy becomes more sparse, and the texture also correlates with star formation rate (Pedersen et al., 2013).

On the other hand, several numerical content descriptors did not show substantial difference between galaxies of different Hubble stages. For instance, Radon features do not show a change between different galaxy types, as well as Tamura textures. The weak ability of Tamura textures to differentiate between galaxy types is that the directionality can be offset by galaxies or arms rotating to the opposite direction. That is different from other texture analysis algorithms such as Haralick, where the texture entropy and energy are independent of the direction.

The experiment was also repeated with the EFIGI galaxy images of the u, g, i, r, and z bands. The resulting phylogenies are displayed by Figure 4.

As the figure shows, the order of the galaxy types somewhat violates the De Vaucouleurs (1959) scheme. The shorter segments between some of the galaxy types show higher similarity deduced by the method, indicating that in some cases the algorithm could not identify the differences between these types. That shows that although the order of the galaxy types deduced by the algorithm is largely in agreement with the sequence described in (De Vaucouleurs, 1959), processing just one band leads to loss of information, and consequently the order and automatic placement of the galaxy types is not as close to the order of De Vaucouleurs (1959) compared to the color images. Color has been identified to correlate with galaxy types (Strateva et al., 2001), and therefore color information can contribute to the ability of the algorithm to analyze the similarities between different types of galaxies. The probability to have the orders of the tree of the u, g, r, i, and z filters by chance is $6.84{\cdot}10^{-5}$, $6.84{\cdot}10^{-5}$, $1.9{\cdot}10^{-6}$, $1.6{\cdot}10^{-3}$, $6.84{\cdot}10^{-5}$,

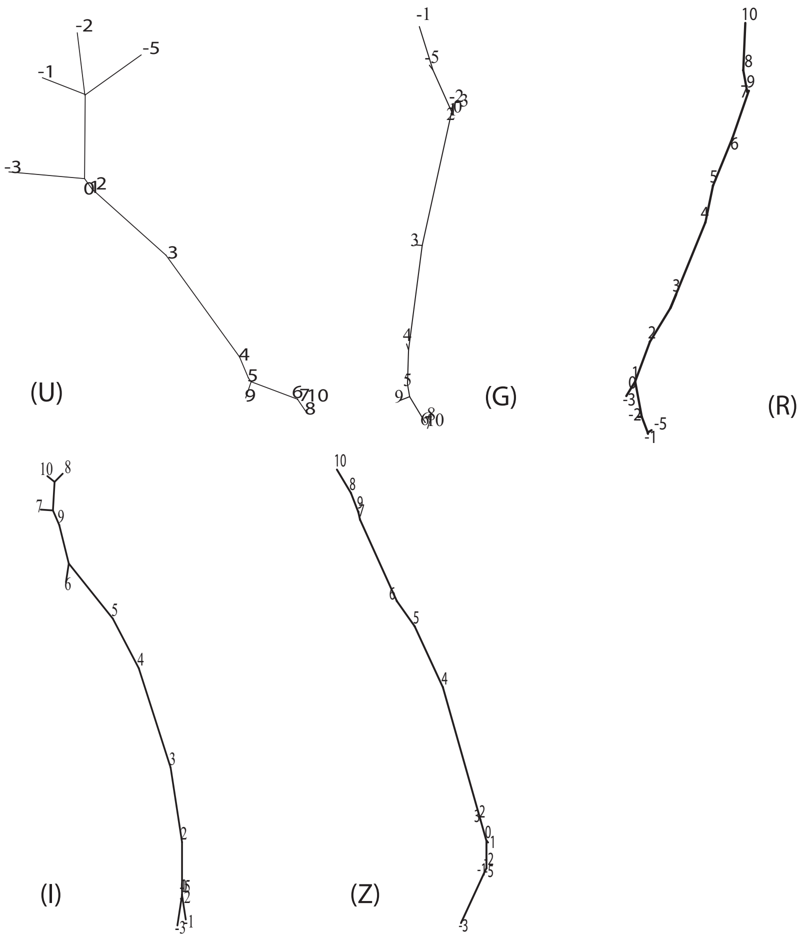

Figure 4: The network of similarities between the galaxy types using the u, g, r, i, and z bands of EFIGI

respectively.

Also, the analysis of the u band shows strong separation between late type galaxies and the other galaxy types, where Sa and Sab are positioned close to the lenticular galaxies. A similar observation can be made with the analysis of galaxies in the g band. The r, i, and z bands show a more even distribution of the types along the main axis, but it is also noticeable that the early types are clustered on one side of the axis, while irregular and Sd galaxies are grouped close to each other on the other side.

Figure 4 also shows that the early type galaxies could not be ordered correctly by the algorithm without using the color images, and that the individual bands or grayscale images did not have sufficient morphological features of these galaxy types that allow the automatic positioning in the same order proposed by De Vaucouleurs (1959).

5 Conclusion

Although galaxy classification cannot be considered a goal in itself (De Vaucouleurs, 1994), it is a key to understanding the physical properties of the past, present, and future universe. Numerous galaxy morphological schemes have been proposed by manual observation and measurement of galaxy morphology and photometry. Here we proposed a computer-based approach to galaxy morphology by using an unsupervised machine learning system that can deduce the visual similarities between sets of images and reconstruct morphological sequences of galaxies. The analysis is performed such that the algorithm determines the network of similarities between the different morphological classes automatically, and without human guidance.

The results show that when using the color EFIGI galaxy images the sequence deduced by the computer is in large agreement with the De Vaucouleurs (1959) scheme, even when using the color images as gray-scale images. When using each band separately the deduced order was in weaker agreement with (De Vaucouleurs, 1959), showing that the composite color images contained more visual information that was used by the algorithm to deduce the order of the morphological types. The saturation and gamma correction applied to the EFIGI color images as described in Section 2 could also affect the way these images were analyzed. Basic statistical analysis shows very low probability of $\sim 1.53 \cdot 10^{-12}$ for having the galaxy types ordered in an ascending or descending order by mere chance.

The color images allowed the algorithm to deduce a sequence that is more consistent with the order of De Vaucouleurs (1959) compared to the sequences produced with each of the individual bands, indicating that the color images contained more information that was used by the algorithm to deduce the order of the morphological typess.

One difference between the De Vaucouleurs (1959) scheme and the network of morphological similarities produced by the algorithm is the S0 galaxies, where the computer algorithm did not find the exact same order identified by De Vaucouleurs (1959). The ability of the computer to deduce a network of similarities that is largely in agreement with manual analysis demonstrates the discovery power of the method, and its potential ability to analyze larger datasets containing a higher number of galaxy classes and identify and profile a possible morphological sequence. That allows quantitative morphological of entire galaxies, rather than the quantification of individual identifiable morphological features (e.g., the number of spiral arms).

While the experiments described in this paper are focused on galaxies in the Hubble sequence, with the increasing importance of digital sky surveys imaging billions of galaxies such as the Large Synoptic Survey Telescope (LSST), automated methods are also important to identify and analyze peculiar galaxies that cannot be associated with a defined morphological stage on the Hubble sequence. The scheme of numerical image content descriptors described in Section 3 has demonstrated its efficacy in detecting peculiar galaxy mergers among millions of galaxies in Sloan Digital Sky Survey, and performing quantitative assessment of these mergers (Shamir and Wallin, 2014). Sky surveys such as LSST will be able to image a much larger number of galaxies, from which peculiar galaxies can be detected. Automatic detection methods such as (Shamir, 2012a; Shamir and Wallin, 2014) can assist in the detection of peculiar galaxies that are not associated with stages on the Hubble sequence, and analysis methods such as the method described here can be used to identify links between a large number of galaxy types.

Source code of the analysis methods used in the experiment are publicly available, as well as the protocol as described in A.

References

Abd Elfattah, M., El-Bendary, N., Abu Elsoud, M. A., Hassanien, A. E., Tolba, M., 2013. An intelligent approach for galaxies images classification. In: Hybrid Intelligent Systems (HIS), 2013 13th International Conference on. IEEE, pp. 167–172.

Abraham, R. G., Van Den Bergh, S., 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.

Almeida, J. S., Aguerri, J., Muñoz-Tuñón, C., De Vicente, A., 2010. Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra. The Astrophysical Journal 714 (1), 487.

Au, K., 2006. Inferring galaxy morphology through texture analysis. Ph.D. thesis, Carnegie Mellon University, ph.D Thesis.

Baillard, A., Bertin, E., de Lapparent, V., Fouqué, P., Arnouts, S., Mellier, Y., Pelló, R., Leborgne, J.-F., Prugniel, P., Makarov, D., et al., 2011. The efigi catalogue of 4458 nearby galaxies with detailed morphology. Astronomy & Astrophysics 532, 74.

Baillard, A., Bertin, E., Mellier, Y., McCracken, H., Géraud, T., Pelló, R., Leborgne, F., Fouqué, P., 2006. Project efigi: Automatic classification of galaxies. In: Astronomical Data Analysis Software and Systems XV. Vol. 351. p. 236.

Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., Brunner, R. J., 2004. Galaxy types in the sloan digital sky survey using supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society 348 (3), 1038–1046.

Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., et al., 2010. Galaxy zoo: reproducing galaxy morphologies via machine learning. Monthly Notices of the Royal Astronomical Society 406 (1), 342–353.

Bishop, C. M., et al., 2006. Pattern recognition and machine learning. Vol. 1. springer New York.

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

Davis, D. R., Hayes, W. B., 2014. Sparcfire: Scalable automated detection of spiral galaxy arm segments. The Astrophysical Journal 790 (2), 87.

de Lapparent, V., Baillard, A., Bertin, E., 2011. The efigi catalogue of 4458 nearby galaxies with morphology ii. statistical properties along the hubble sequence. Astronomy & Astrophysics 532, 236–239.

De Vaucouleurs, G., 1959. Classification and morphology of external galaxies. In: Astrophysik IV: Sternsysteme/Astrophysics IV: Stellar Systems. Springer, pp. 275–310.

De Vaucouleurs, G., 1994. Global physical parameters of galaxies. In Proceedings: Quantifying galaxy morphology at high redshift. Space Telescope Science Institute, Baltimore, MD.

De Vaucouleurs, G., De Vaucouleurs, A., Corwin Jr, H., Buta, R., Paturel, G., Fouque, P., 1992. Third reference catalogue of bright galaxies (rc3). VizieR Online Data Catalog 7137, 0.

Dieleman, S., Willett, K. W., Dambre, J., 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. arXiv preprint arXiv:1503.07077.

Felsenstein, J., 1993. {PHYLIP}: phylogenetic inference package, version 3.5 c.

Gabor, D., 1946. Theory of communication. part 1: The analysis of information. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering 93 (26), 429–441.

Gradsteyn, I., Ryzhik, I. M., 1994. Alan jeffrey: Table of integrals, series and products.

Grigorescu, S. E., Petkov, N., Kruizinga, P., 2002. Comparison of texture features based on gabor filters. Image Processing, IEEE Transactions on 11 (10), 1160–1167.

Hadjidemetriou, E., Grossberg, M. D., Nayar, S. K., 2001. Spatial information in multiresolution histograms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. Vol. 1. IEEE, pp. I–702.

Haralick, R. M., Shanmugam, K., Dinstein, I. H., 1973. Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics (6), 610–621.

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

Huertas-Company, M., Aguerri, J., Bernardi, M., Mei, S., Almeida, J. S., 2010. Revisiting the hubble sequence in the sdss dr7 spectroscopic sample: a publicly available bayesian automated classification. arXiv preprint arXiv:1010.3018.

Humason, M. L., Mayall, N. U., Sandage, A. R., 1956. Redshifts and magnitudes of extragalactic nebulae. The Astronomical Journal 61, 97–162.

Kormendy, J., Bender, R., 1996. A proposed revision of the hubble sequence for elliptical galaxies. The Astrophysical Journal Letters 464 (2), L119.

Kormendy, J., Djorgovski, S., 1989. Surface photometry and the structure of elliptical galaxies. Annual review of astronomy and astrophysics 27, 235–277.

Kuhner, M. K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11 (3), 459–468.

Kuminski, E., George, J., Wallin, J., Shamir, L., 2014. Combining human and machine learning for morphological analysis of galaxy images. Publications of the Astronomical Society of the Pacific 126 (944), 959–967.

Lekshmi, S., Revathy, K., Nayar, S. P., 2003. Galaxy classification using fractal signature. Astronomy & Astrophysics 405 (3), 1163–1167.

Lim, J. S., 1990. Two-dimensional signal and image processing. Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p. 1.

Manning, S., Shamir, L., 2014. Chloe: A software tool for automatic novelty detection in microscopy image datasets. Journal of Open Research Software 2 (1), e25.

Meneses Cuadros, E., Plata Gómez, A., Vera-Villamizar, N., 2009. Classification of galaxies using automatic learning algorithms: Sequential solution and parallel design. In: Revista Mexicana de Astronomia y Astrofisica Conference Series. Vol. 35. p. 311.

Morgan, W., 1958. A preliminary classification of the forms of galaxies according to their stellar population. Publications of the Astronomical Society of the Pacific, 364–391.

Morgan, W., Lesh, J. R., 1965. The supergiant galaxies. The Astrophysical Journal 142, 1364.

Morgan, W., Mayall, N., 1957. A spectral classification of galaxies. Publications of the Astronomical Society of the Pacific, 291–303.

Naim, A., Lahav, O., Buta, R., Corwin, H., De Vaucouleurs, G., Dressler, A., Huchra, J., Van den Bergh, S., Raychaudhury, S., Sodre, L., et al., 1995. A comparative study of morphological classifications of apm galaxies. Monthly Notices of the Royal Astronomical Society 274 (4), 1107–1125.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D. M., Goldberg, I. G., 2008. Wnd-charm: Multi-purpose image classification using compound image transforms. Pattern Recognition Letters 29 (11), 1684–1693.

Pedersen, K. S., Stensbo-Smidt, K., Zirm, A., Igel, C., 2013. Shape index descriptors applied to texture-based galaxy analysis. In: International Conference on Computer Vision. IEEE, pp. 2440–2447.

Peng, C. Y., Ho, L. C., Impey, C. D., Rix, H.-W., 2002. Detailed structural decomposition of galaxy images. The Astronomical Journal 124 (1), 266.

Sandage, A., 1961. The Hubble atlas of galaxies. Vol. 618. Carnegie Institution of Washington Washington.

Sandage, A., Freeman, K. C., Stokes, N., 1970. The intrinsic flattening of e, so, and spiral galaxies as related to galaxy formation and evolution. The Astrophysical Journal 160, 831.

Shamir, L., 2006. Human perception-based color segmentation using fuzzy logic. In: International Conference on Image Processing, Computer Vision, & Pattern Recognition. pp. 496–502.

Shamir, L., 2008. Evaluation of face datasets as tools for assessing the performance of face recognition methods. International Journal of Computer Vision 79 (3), 225–230.

Shamir, L., 2009. Automatic morphological classification of galaxy images. Monthly Notices of the Royal Astronomical Society 399 (3), 1367–1372.

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

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

Shamir, L., 2012b. Computer analysis reveals similarities between the artistic styles of van gogh and pollock. Leonardo 45 (2), 149–154.

Shamir, L., Holincheck, A., Wallin, J., 2013a. Automatic quantitative morphological analysis of interacting galaxies. Astronomy and Computing 2, 67–73.

Shamir, L., Ling, S. M., Scott, W., Hochberg, M., Ferrucci, L., Goldberg, I. G., 2009a. Early detection of radiographic knee osteoarthritis using computer-aided analysis. Osteoarthritis and Cartilage 17 (10), 1307–1312.

Shamir, L., Ling, S. M., Scott, W. W., Bos, A., Orlov, N., Macura, T. J., Eckley, D. M., Ferrucci, L., Goldberg, I. G., 2009b. Knee x-ray image analysis method for automated detection of osteoarthritis. IEEE Transactions on Biomedical Engineering 56 (2), 407–415.

Shamir, L., Macura, T., Orlov, N., Eckley, D. M., Goldberg, I. G., 2010a. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. ACM Transactions on Applied Perception 7 (2), 8.

Shamir, L., Manning, S., Wallin, J., 2014a. Chloe: A tool for automatic detection of peculiar galaxies. Astrophysics Source Code Library, ascl:1409.008.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston, J., Goldberg, I., 2013b. Wnd-charm: Multipurpose image classifier. Astrophysics Source Code Library, ascl:1312.002.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston, J., Goldberg, I. G., 2008a. Wndchrm an open source utility for biological image analysis. Source code for biology and medicine 3, 13.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T. J., Goldberg, I. G., 2008b. Iicbu 2008: a proposed benchmark suite for biological image analysis. Medical & Biological Engineering & Computing 46 (9), 943–947.

Shamir, L., Orlov, N., Goldberg, I. G., 2009c. Evaluation of the informativeness of multi-order image transforms. In: International Conference on Image Processing, Computer Vision and Pattern Recognition. pp. 37–42.

Shamir, L., Rahimi, S., Ferrucci, L., Goldberg, I., et al., 2010b. Progression analysis and stage discovery in continuous physiological processes using image computing. EURASIP Journal on Bioinformatics and Systems Biology 2010.

Shamir, L., Tarakhovsky, J. A., 2012. Computer analysis of art. Journal on Computing and Cultural Heritage 5 (2), 7.

Shamir, L., 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), 3528–3537.

Shamir, L., Wallin, J. F., Allen, A., Berriman, B., Teuben, P., Nemiroff, R. J., Mink, J., Hanisch, R. J., DuPrie, K., 2013c. Practices in source code sharing in astrophysics. Astronomy and Computing 1, 54–58.

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A. M., Tyack, P., Samarra, F., Miller, P., Wallin, J., 2014b. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. The Journal of the Acoustical Society of America 135 (2), 953–962.

Simard, L., 1999. Photometric redshifts and the luminosity-size relation of galaxies to z= 1. 1. In: Photometric Redshifts and the Detection of High Redshift Galaxies. Vol. 191. p. 325.

Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., McConnachie, A. W., 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.

Strateva, I., Ivezić, Ž., Knapp, G. R., Narayanan, V. K., Strauss, M. A., Gunn, J. E., Lupton, R. H., Schlegel, D., Bahcall, N. A., Brinkmann, J., et al., 2001. Color separation of galaxy types in the sloan digital sky survey imaging data. The Astronomical Journal 122 (4), 1861.

Tamura, H., Mori, S., Yamawaki, T., 1978. Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics 8 (6), 460–473.

Teague, M. R., 1980. Image analysis via the general theory of moments. Journal of the Optical Society of America 70 (8), 920–930.

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

Van Den Bergh, S., 1976. A new classification system for galaxies. The Astrophysical Journal 206, 883–887.

Wu, C.-M., Chen, Y.-C., Hsieh, K.-S., 1992. Texture features for classification of ultrasonic liver images. IEEE Transactions on Medical Imaging 11 (2), 141–152.

A Protocol

All software tools used to produce the results are open source, making it easier to replicate the results or analyze other datasets (Shamir et al., 2013c). Source code for the computer analysis method (Shamir et al., 2008a) is available at the Astrophysics Source Code Library (Shamir et al., 2013b) or at http://vfacstaff.ltu.edu/lshamir/downloads/ImageClassifier, as well as its dependency libraries (Shamir et al., 2013b). It also requires the installation of the open source PHYLIP package, available at http://evolution.genetics.washington.edu/phylip.html. The experiments also require computational resources that can process the EFIGI catalog. The experiment in this paper was done with a 16-core Intel Core-i7 machine and 32GB of RAM, and took about three days to complete.

To replicate the results, the following steps are required:

1. Download the EFIGI catalog from http://www.astromatic.net/projects/efigi

2. Convert the color FITS images (or PNG images) to TIF format by using ImageMagick. A batch conversion can be done by the following command line: find /path/to/efigi -name “*.FITS” -exec convert {} {}.tif \;

3. Separate the images into folders such that the name of each folder is the T number, and its content is the galaxy images of that T number as annotated by EFIGI.

4. Compute the image features by running the command line: ./wndchrm train -mlc /path/to/efigi_root_folder /path/to/efigi_root_folder/efigi.fit
That step might take several days to complete with a single core, but the response time can be shorter by running several instances of the process. The process should not be stopped to avoid the creation of empty .sig files. In case the process stopped before completion, the following command line should be used before running it again: find /path/to/efigi -name “*.sig” -exec rm {} \;

5. The phylogeny can be created by running the following command line: ./wndchrm test -f0.15 -i#140 -j12 -n20 -w -p/path/to/phylip /path/to/efigi_root_folder/efigi.fit /path/to/efigi_root_folder/efigi.html
When done, a .ps file should be created in the folder “/path/to/efigi_root_folder”.

6. To process the grayscale images step 4 should be replaced with the command line: ./wndchrm train -ml -S100:25 /path/to/efigi_root_folder /path/to/efigi_root_folder/efigi.fit

The experiments were performed in Linux (Fedora) environment. For further information or assistance please contact the authors.