Galaxy Image Classification using Non-Negative Matrix Factorization

I.M.Selim
National Research Institute of
Astronomy and Geophysics
(NRIA)

Arabi E. Keshk
Faculty of computer and
information,
Menofya University

Bassant M.El Shourbugy
Computer Science Department
Higher Technological Institute
(HTI)

ABSTRACT

In modern astronomy with the advent of astronomical imaging technology developments and the increased capacity of digital storage, lead to the production of photographic atlases of data which need to be processed autonomously. Galaxies morphology is an important topic to understand questions concerning the evolution and formation of galaxies and their content. In this work, morphological classification of galaxies is presented using a new method based on Non-Negative matrix factorization for images of galaxies in the Zsolt frei Catalog. The algorithm is trained using manually classified images of elliptical, spiral and lenticular galaxies. Experimental results show that galaxy images from Zsolt Frei catalog can be classified automatically with an accuracy of 93 percent compared to classifications carried out by other authors and manually classified.

General Terms

Pattern recognition; image classification.

Keywords

Galaxy classification; Hubble's classification scheme; De Vaucouleurs; Non-negative Matrix Factorization (NMF).

1. INTRODUCTION

Galaxies are mysterious beautiful creations with all its wide variety of appearances, started by the creation of the universe 13.7 billion years ago which is known by the Big Bang theory which lead to all the different types of galaxies [1]. Studying the types and the properties of galaxies is important as it offers important clues about the origin and developments in the universe. Large catalogs were used by astronomers to test theories and study the underlying physics of the universe. Sloan Digital Sky Survey (SDSS) is one of the most successful modern data collection projects in astronomy, it use a dedicated 2.5-m wide-angle optical telescope for multi-filtering imaging and spectroscopic surveys, its data collection started in 2000 and it covered most of the sky area in its final data release [2]. A major subset of the survey is catalogs of galaxy images. This has the potential to be a valuable source of data to astronomers. [3].

In the past, galaxies used to be categorized manually into categories based on their visually observed characteristics, but as the modern sky surveys provide millions of galaxies images, it's found that applying classification algorithms will help in solving this problem.

In the past few years, advancements in computational tools and algorithms have started to allow automatic analysis of galaxy morphology. Previous work of galaxies classification helps in understanding the processes that created galaxies that share similar structures. Several machine learning algorithms have been used in the few past years to automate the classification process. [4], used supervised ANN to classify galaxies, Difference Boosting Neural Networks was able to learn 98.3% of the galaxies correctly and identify 89.9% of galaxy images in a test set; their challenge was to develop a supervised classifier capable of sorting galaxies into subclasses using manually threshold images. [5], Presented an experimental study of machine learning and image analysis for performing automated morphological galaxy classification. They used a neural network, and a locally weighted regression method, and implemented homogeneous ensembles of classifiers. The ensemble of neural networks was created using the bagging ensemble method, and manipulation of input features was used to create the ensemble of locally weighed regression. Principle Component Analysis was used to reduce the dimensionality of the data, and to extract relevant information in the images. [6], compared and evaluated the behavior of 10 artificial neural networks based classifiers based on selected sets of galaxy features derived from image analysis and principle component analysis features (PCA). The support vector machine (SVM) based classifier provides the best results about 99.5%. [7], presented an image analysis supervised learning algorithm that can automatically classify galaxy images. He found that galaxy images from Galaxy Zoo can be classified automatically to spiral, elliptical and edge on galaxies with an accuracy of ~90% compared to classifications carried out manually. [8], Proposed an algorithm that classifies galaxy images using invariant moment's features he showed that images can be classified with accuracy 90% compared to the human's visual classification system.

Non-negative matrix factorization (NMF), also known as non-negative matrix approximation [9][10] is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements. (NMF) is one of the recently arisen dimensionality reduction methods. Unlike other methods like Principle component analysis (PCA), (NMF) is based on non-negative constraints, which allow learning parts from objects [11]. (NMF) was first introduced by Paatero and Tapper in [12] but it has gained popularity by the work of Lee and Seung [13]. They argue that the non-negative is important in human perceptions and also give simple algorithms for finding a non-negative representation for non-negative data.

The paper is organized as follows: Section 2 gives a brief introduction to Hubble's galaxy classification scheme, De Voucleurs Galaxy morphological classification and the previous work of other authors in the field of Galaxy classification automation. In section 3 we describe the system architecture and the used algorithm. Section 4 gives a brief explanation for the database of images used and the experimental results. Finally, in section 5 (last section) we presented the conclusion and future work.

2. GALAXY MORPHOLOGICAL CLASSIFICATION

2.1 Hubble's Classification Scheme

Galaxies show a vast range of forms, Astronomers for many decades have been trying to find the underlying order in the diverse properties of galaxies [6]. They have found that galaxies can be categorized based on their physical appearances. One of the most common classification schemes is the system devised by Sir Edwin Hubble in 1936, It is often known as the “Hubble tuning-fork", Hubble’s scheme divides galaxies into three classes based on their visual appearance: (i) Elliptical; (ii)Spiral; (iii) Barred spiral; and (iv)Irregular [6]. As shown in figure (1).

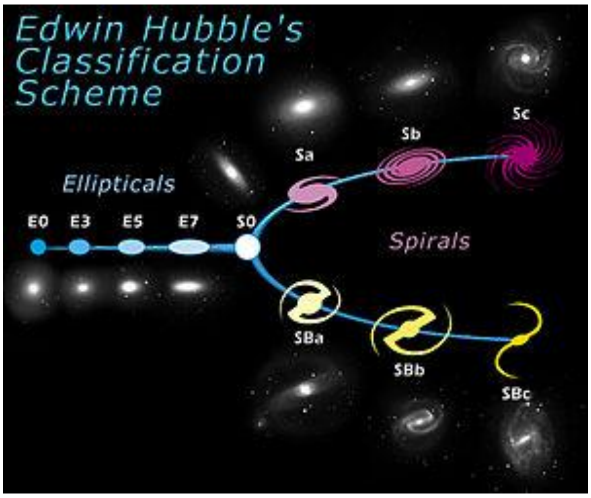

Fig. 1. Hubble's Classification Scheme

- Elliptical galaxies: E0, E3, E5, and E7; have almost no discernible structure, they have smooth light distributions and appear as ellipses in images. They are denoted by the letter E, followed by an integer n representing their degree of ellipticity on the sky.
- Spiral galaxies: S0, Sa, Sb, Sc, and Sd; are the most common type of galaxies, they look like a flattened disk, with stars forming a spiral arms winding towards a central concentration of stars known as the bulge, spirals are given the symbol "S". Almost half of all spirals are also observed to have a bright line, or a bar, running through them extending from the central bulge. These barred spirals are given the symbol "S.B.".
- Lenticular galaxies: SBa, SBb, and SBc; are the intermediate between spiral and elliptical, they are labeled S0 or SB0, it consists of a very bright bulge with no surrounded arms and that what makes it difference from an E0 galaxy [6].
- Irregular galaxies: Im, and Ibm, some galaxies don't have neither a spiral structure nor a nuclear bulge, they appear as a random collection of stars with no obvious order.

2.2 De Vaucouleurs Classification

The De Vaucouleurs system is considered a revised extension to Hubble's system, it was developed by Gérard de Vaucouleurs in 1959 [14], this system provided a finer level of discrimination than Hubble's system that was considered the basics of galaxy classification since he was the first to provide an order of galaxies that represents those galaxies can be morphologically different [15].

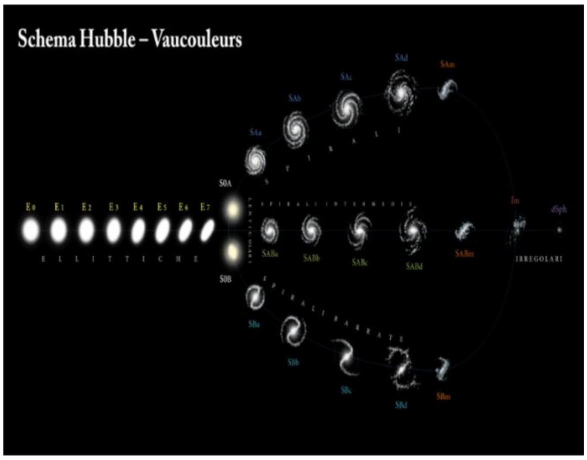

Fig.2. De Vaucleurs system.

De Vaucouleurs system, as shown in figure (2), is considered a compliment to Hubble's scheme, He introduced a more elaborate classification system for spiral galaxies, based on the three basic morphological characteristics:

- Bars, galaxies division is based on the presence or absence of a nuclear bar. De Vaucouleurs introduced the notation SA to define spiral galaxies without bars, an intermediate class SAB was also introduced to denote weakly barred spirals [16], he also used the notation S0 to describe Lenticular galaxies that are impossible to tell whether they have a bar or not, and for barred Lenticulars he used the notion SB0 while he used the notion SAB for unbarred Lenticulars.
- Rings. Galaxies with ring-like structures (denoted ‘(r)’) and those without rings (denoted ‘(s)’). While ‘transition’ galaxies are given the symbol (rs).[16]
- Spiral arms. Hubble's scheme categories spiral galaxies into classes based on the tightness of their spiral arms, De vaucouleurs extended the spiral classes by adding several additional classes, these additional classes were classified in Hubble's scheme as Irregulars Irr, De Vaucouleur used the notions: Sd (SBd), Sm (SBm), Im. In addition, the Sd class contains some galaxies from Hubble’s Sc class.

3. THE PROPOSED METHOD

3.1 System Architecture

The architecture of the system goes as follows:

1. Image pre-processing, in which images are getting resized, scaled, rotated and centered, getting it ready for the second phase.
2. Feature extraction, in which morphological features are extracted from the images, this phase is generally used to minimize the dimensionality of galaxy data, the method that has been proposed to construct features is a good

reconstruction of images. Therefore, each feature can be considered again as an image. Together with the participation of each feature in an image, one can establish the composition of every image in a very comprehensible way. The NMF method has been proposed as a novel subspace method in order to obtain a parts-based representation of objects by imposing non-negative constraints.

3. Classification procedure, in which the computer is trained using a set of data that are based on the classification provided by human experts.

An implementation of Non-negative Matrix Factorization is described for purposes of galaxy classification; the algorithm consists of two main steps: The first is the training step in which each training sample is normalized. The second main step is the test step, where the test set is all images. In the experiment testing was done using the gray-scale, images and normalized all images with respect to intensity for intensity with the same standard deviation. Normalization is done to ensure that the order will be determined by the shape, with no impact of color or brightness see Figure (3).

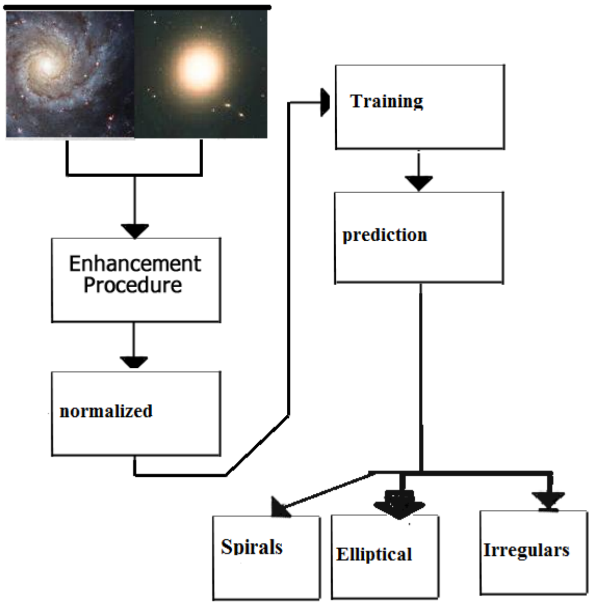

Fig. 3. Non-negative Matrix Factorization Galaxy Classification Scheme.

3.2 Non –negative Matrix Factorization

Non-negative matrix factorization is a linear dimensionality reduction technique which is useful in handling nonnegative data [13]. It allows only non-subtractive combinations of nonnegative basis vectors, leading to (possibly) a parts-based representation. And that`s what makes it distinguished from other methods as Principle Component Analysis (PCA) by its non-negativity constraints. These constraints lead to part-based representation because they allow only additive, not subtractive combinations of the original data [13] [17]. Given an initial database of matrix *V* with dimensions n × m where m is the number of examples in the dataset. Matrix *V* is factorized into two matrices W with dimensions n × r and *H* with dimensions r × m, [18] usually *r* is chosen to be smaller than *n* or *m* so that the two matrices *W* and *H* becomes less than the original matrix *V*, this result a compressed version of the original data matrix [19].The algorithm used is the NMF standard algorithm.

$$V \approx WH \qquad (1)$$

A common way to find W and H is by minimizing the Euclidean distance between V and WH [18].

$$\min_{H,W} f(W,H) = \frac{1}{2}||V - WH||_5^2$$

$$\text{(2)} \qquad \text{Subject to } \forall i, j\ , W_{ij}\ , H_{ij}\ \geq 0$$

Where $||.||_F$ is Frobenius norm, the interpretation of W and **H** is different based on the application.

Where W, H ≥ 0 means that all elements of matrices W and H are non-negative (20).

The algorithm consists of two main steps: The first is the training step in which each training sample is normalized to make all images have the same scale, by decomposes these sets into basis matrix W and coefficient matrix H. The second main step is the test or prediction step, where the test set S is normalized also. The coefficient matrix **A** of the S is computed based on basis matrix W, where the basis matrix W contains all information about different galaxy types (i.e spiral, elliptical, etc..) and this can lead to separate the images belong to the same type (like blind source separation). To predict the class of an unknown sample $\boldsymbol{S_i}$, we used a MAX rule that selects the maximum coefficient (the coefficient vector of $\boldsymbol{A}$), and then assign the class label of the corresponding training sample to this new sample.

```
Algorithm. NMF Classifier

Input: $V$: training set
         $r$: cluster numbers
         $S$: $P$ unknown samples without labels
Output: $P$: predicted class labels of the p unknown samples
Training Step:
    1. Normalize training set to have $l_2$-norm
    2. Solve the NMF optimization problem:
       [$W, H$] to ($V, r$)
Test Step:
    1. Normalize test set to have normalized
    2. Solve the NMF optimization problem:
       $\min_{A} f(W, A) = \frac{1}{2}||S - WA||_F^2$
    3. Predict the class label:
       $p_i = MAX(a_i)$
    4. Return $P$.
```

4. EXPERIMENTAL RESULTS

The images used in this paper were obtained from the Zsolt frei Catalog [21]. This Galaxy Catalog is a collection of digital images of 113 nearby galaxies. Images taken in several passbands and a color-composite image are included for each galaxy [21]. Images of 31 galaxies were taken with the 1.5-meter telescope at the Palomar Observatory in 1991; images of the other 82 galaxies images were taken with the 1.1-meter telescope at the Lowell Observatory in 1989. At Palomar, a camera with an 800x800 TI CCD was used, at Lowell, the camera had an RCA 512x320 CCD. Palomar images are available in three passbands of the Thuan-Gunn system: g, r, and i. [22] Lowell images are in two passbands (J and R) of the filter system developed by [23] Gullixon et al. The selected data set of Zsolt Frei catalog has the following properties (i) high resolution (ii) good quality (iii) careful calibration. Which doesn't require high preprocessing phases such as gamma correction or histogram equalization [6]. The database was used before in many galaxy classification methods using other techniques such as Artificial Neural Network (ANN). References [4] [8] used images from the Zsolt Frei catalog in their classification techniques as denoted in the introduction section. The iteration and all the experiments were performed on a Windows 7 Ultimate 64-bit operating system; processor Intel Core i5 760 running at 2.81 GHz; 4 GB of RAM and code was implemented in MATLAB with runtime about few seconds less than one minute.

The obtained accuracy is satisfactory and is comparable to that of a specialist. On a training set of 113 images, a 93% of correct classification was obtained from three Hubble types (see table 1), using a set of 20 images for testing and obtaining 105 accurate results, 3% inaccurate images when trying to discriminate spiral images and 4% inaccurate in the case of discriminating lenticular images. It is also noticeable that the S0 galaxy types are different. This result indicating that the percent value different across different galaxy morphological types. Some images were not classified well while others were well classified, see figure (4), (5). Thus, It can be assumed that the classification accuracy can be improved when using datasets with less noise and solving the overlapping problem.

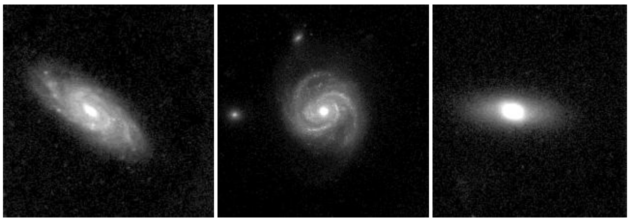

Fig. 4. A sample of images from our data that were well classified.

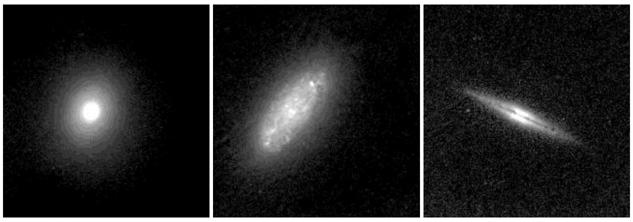

Fig. 5. A sample of images that were not good in classification.

Table 1; give the percentage of different types of galaxies, in our data that are put into the different classes by the Nonnegative Matrix Factorization.

Average Accuracy	Spirals	Ellipticals	Irregulars
	Accuracy	Accuracy	Accuracy
92.76	0.92.1	0.91.2	0.95

5. CONCLUSIONS

In this study, we have proposed a computer-based approach to classification galaxies morphology by using the supervised machine learning system based on Non-Negative matrix factorization algorithm, that can derive the automatically classify images of the spiral, elliptical and lenticular galaxies similarities visual classification of galaxies images. The classification of elliptical type and spiral type classifications of galaxies in our sample using Non-Negative matrix factorization algorithm match those by the human eye to better than 97 percent. The analysis is performed such that the algorithm determines the morphology types of galaxies from different morphological classes automatically, and without human guidance. The results show that when using only two galaxies type from Hubble sequence deduced by the computer, there is in large agreement with the visual classification.

The algorithm was able to reproduce the human classifications for the rest of the objects to better than 90 percent in three morphological classes. So the most important parameters are those which relate directly to galaxy evolution. We assume that it will be reasonable to be used for other problems in the morphological analysis of celestial objects.

Future research could focus on improving the classification results and the runtime of the algorithms.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the invaluable contribution of Dr. Mohamed Abd El Aziz, Department of mathematics, Faculty of science, Zagazig University for his helpful discussion comments.

7. REFERENCES

[1] Wollack ،E. J. "Cosmology: The Study of the Universe".Universe 101: Big Bang Theory. NASA. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73, December.

[2] Gunn.et.al,"The 2.5 m Telescope of the Sloan Digital Sky Survey", 2006.

[3] K.N Abazajian. "The seventh data release of the Sloan digital sky survey". Astrophysical journal supplement series, 2009.

[4] Goderya, S., Andreasen, J. D., & Philip, N. S." Advances in Automated Algorithms For Morphological Classification of Galaxies Based on Shape Features", Astronomical Data Analysis Software and Systems (ADASS) XIII, Vol. 314. San Francisco: Astronomical Society of the Pacific, 2004.

[5] J.Calleja and O.Fuentes."Machine learning and image analysis for morphological galaxy classification". Monthly Notices of the Royal Astronomical Society, Volume 349, Issue 1, pp. 87-93, 2004.

[6] M.M.Ata, M. A. Mohamed, H.K. El-Minir, and A.I. Abd-El-Fatah., "Automated classification techniques of galaxies using artificial neural networks based classifiers", IEEE, 2009.

[7] Shamir L, "Automatic morphological classification of galaxy images", Instrumentation and Methods for Astrophysics, 2009.

[8] M.Abdelfattah, M Abu Elsoud, A.E Hassanein, "Automated classification of galaxies using invariant moments", Proceedings of the 4th international conference on Future Generation Information Technology, 2015.

[9] Berman A.and Plemmons R.J." Nonnegative matrices in the mathematical sciences". Siam, 1994.

[10] Tandon, Rashish; Suvrit Sra "Sparse nonnegative matrix approximation: new formulations and algorithms",2010.

[11] Oleg G. Okun, "Non-negative matrix factorization and classifiers: Experimental study", Proc. of the Fourth IASTED International Conference on Visualization,

Imaging, and Image Processing (VIIP 2004), Marbella, Spain, 2004.

[12] Paatero P. and Tapper U. " Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics", 5(1):111–126, 1994.

[13] Lee, D. D., & Seung, H. S., "Learning the parts of objects by non-negative matrix factorization. Nature", 401, 788–791, 1999.

[14] De Vaucouleurs, G. "Classification and Morphology of External Galaxies". Handbuch der Physik 53: 275. Bibcode:1959HDP....53..275D. 1959.

[15] Binney, J., Merrifield, M., "Galactic Astronomy". Princeton: Princeton University Press, 1998

[16] De Vaucouleurs, Gérard."Revised Classification of 1500 Bright Galaxies", Astrophysical Journal Supplement 8:31.Bibcode: 1963ApJS….8…31D. April 1963.

[17] Paatero, P & Tapper, U. "Least squares formulation of robust non-negative factor analysis", Chemometr. Intell. Lab. 37, 23- 35, 1997.

[18] Chih-Jen Lin, "On the Convergence of Multiplicative Update Algorithms for Non-negative Matrix Factorization", IEE Transactions on neural networks, vol,18, No.6, November 2007

[19] E. Benetos, M. Kotti, C. Kotropoulos. "Musical instrument classification using non-negative matrix factorization algorithms", May 2006

[20] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,”Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference.

[21] Z.Frei, P. Guhathakurta, J. E. Gunn and J. Anthony Tyson, "A catalog of digital images of 113 nearby galaxies", Astronomical Journal in January 1996.

[22] G.W.Zack, W.E. Rogers, and S.A. Latt, "Automatic Measurement of Sister Chromatid Exchange Frequency", Journal of Histochemistry and Cytochemistry Vol. 25, No. 7, pp.741-753,1997.

[23] GULLIXSON C.A.; BOESHAAR P.C.; TYSON J.A.; SEITZER P," The BjRI photometric system ", *1995ApJS...99..281G - Astrophys. J., Suppl. Ser., 99, 281-293 (1995) - 01.01.86 01.01.86 July 1995.*