Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference

A Hierarchical Model for Morphological Galaxy Classification

Maribel Marin and **L. Enrique Sucar** and **Jesus A. Gonzalez** and **Raquel Diaz**
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Tonantzinlta, Puebla, Mexico
{mmarinc, esucar, jagonzalez, raqueld }@inaoep.mx

Abstract

We propose a new method for the morphological galaxy classification which incorporates two main contributions: (i) the generation of artificial images of galaxies through geometric transformations to be used as additional examples in the training phase, (ii) the use of a novel hierarchical classifier for hierarchical galaxy classification. An additional classifier distinguishes galaxies from stars based on geometrical moments. The proposed method was tested with two different astronomical databases. The results found show that the hierarchical classification method has a higher performance than flat classification, and that the use of artificial examples and oversampling provide a significant improvement in performance.

Introduction

Galaxy classification is important for different reasons. First, we are able to produce large catalogs used to statistically analyze those observation. Second, it allows discovering the underlying physics as described in Lahav's work (Lahav 1996). There are two ways to approach the galaxies classification problem: morphological and spectral. Morphology based classification describes the galaxies appearance, while Spectral galaxies classification considers their stellar composition.

Different methods for morphological galaxies classification have been proposed such as the use of classification ensembles (Bazell and Aha 2001). This method uses a set of classifiers for training and then it combines their final predictions by a voting or an average technique. Some other techniques, such as *Resampling* or *SMOTE* are used to deal with the class imbalance problem (De la Calleja et al. 2010). However, two main problems remain. On one hand, the performance decreases as the number of different classes is increased, in particular if sub-classes are considered. On the other hand, the class imbalance issue still remains, and this also get worse as we consider more classes.

In this work we propose a methodology for morphological galaxy classification that includes two main contributions. The first one is the generation of artificial images of galaxies through geometric transformations to be used as additional examples in the training phase. The second one is the use of a novel hierarchical classifier, to our knowledge this is the first time that a hierarchical scheme has been used for galaxy classification. Additionally, we propose a method that performs a feature extraction process in which the galaxy nucleus and the rest of the galaxy body are considered separately. This step provides stronger features for the morphological classification task. We also separate the galaxies from other astronomical objects in the images, with an additional classifier that distinguishes galaxies from stars based on geometrical moments.

We evaluate our method using a database created from the astronomical plates of the Schmidt camera situated at the National Institute for Astrophysics, Optics, and Electronics (INAOE). We also analyzed the database described in (Bazell 2000). We used *Naive Bayes* and *Random Forest* as base classifiers. We compared our method with the one proposed by (Bazell and Aha 2001). The experimental evaluation shows that our method is significantly better than a flat classification approach as well as alternative techniques as the one described in (Bazell and Aha 2001).

Galaxies Morphological Classification

There are galaxies with different morphologies. In 1926, Edwin Hubble created a galaxy classification method based on their shape (Karttunen et al. 2007). In Hubble's classification scheme there are three main types of galaxies: spiral, elliptical, and irregular; as we can see in Figure 1.

Hierarchical Classification

Hierarchical classification can be seen as particular type of multidimensional classification in which the output of a classification algorithm is generated by following a taxonomy. A classification taxonomy may have either a directed acyclic graph (DAG) or a tree structure.

The main hierarchical classification methods are:

- *Flat Hierarchical Classification*. This is the simplest hierarchical classification method (Xiao et al. 2007). It completely ignores the classes hierarchy and classifies only using the leaf nodes.
- *Local Hierarchical Classification*. In this case, the hierarchy is considered by the use of local information (Koller and Sahami 1997), which can be used in different ways.

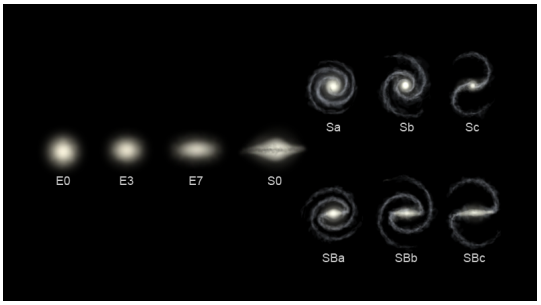

Figure 1: Hubble's diapason diagram. This figure shows the three main galaxies types and their respective subclasses. E0-E7 are elliptical galaxies. Sa, Sb, and Sc are normal spiral galaxies. SBa, SBb, and SBc are barred spiral galaxies. Finally, S0 are lenticular galaxies.

These approaches are differentiated by the way in which they use local information and how they build the classifiers. Then there are three ways to use local information:

- Local Classifier per Node. This is the most used approach. It consists in the creation of a binary classifier for each node in the hierarchy (except by the root node).
- Local Classifier by Parent Node. In this case, the classifiers work only in the parent nodes of the hierarchy, there are no classifiers in leaf nodes.
- Local Classifier per Level. In this approach we build a multi-class classifier for each level of the hierarchy.

- *Global hierarchical Classification (Big-Bang)*. In this case, a global classification model is created (Freitas and de Carvalho 2007). It has the advantage that the size of the whole global classification model is usually smaller (comparing with the total size of local models). Its disadvantage is that when the number of classes increases or decreases, the model has to be rebuilt.

The most commonly used techniques are the local classifiers in a top-down scheme; however these have the problem that if a local classifier provides an incorrect classification this is propagated down the hierarchy. Recently an alternative hybrid scheme has been proposed, which is described below.

Multidimensional Hierarchical Classification

Hierarchical classification is a variant of multidimensional classification, with the difference that classes are organized in a hierarchy. In this case, it is possible to consider a multidimensional classification approach (Julio Hernandez 2013) as a hierarchical classification method. Such a method has as its objective finding the best path to classify a new sample, by combining the probabilities of the classifiers in that path. Each classifier in the path provides a probability that indicates the confidence with which the sample should be classified. The probabilities obtained in a path are then combined to provide a final classification: the classes corresponding to the path with the highest probability.

There are three alternatives for finding the *best* path:

- **Descendent Probabilities Order (DOP)**. Classes predicted by local classifiers are ordered in a descendent way according to their probability value. According to this order, we look for the first consistent subset of classes, that is, a path from the root to one of the leaves. Then, this path is output as the global prediction.
- **Sum of Probabilities (SP)**. In this case the probabilities of the most probable class of the local classifiers in each path of the hierarchy are added. Thus, the global prediction is the path with the highest sum.
- **Product of Probabilities PP)**. In this case the probabilities of the most probable class of the local classifiers in each path of the hierarchy are multiplied. Thus, the global prediction is the path with the highest product.

An example of this hierarchical classification method can be seen in Figure 2

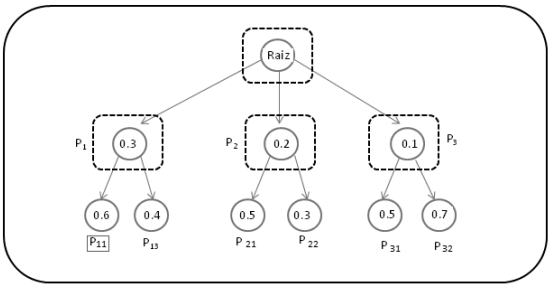

Figure 2: Multidimensional Hierarchical Classification. Each node (circle) represents each of the classes of the hierarchy. Dotted rectangles represent the classifiers and the numbers inside the nodes are the probabilities of each class. In this case the predicted classes for the three alternatives are P_1 and P_{11}, as this path has the highest product,**PP=**0.18, and sum **SP=**0.9. For **DPO**, the probabilities have the following order $P_{32}, \mathbf{P_{11}}, P_{21}, P_{31}, P_{13}, P_{22}, \mathbf{P_1}, P_2, P_3$, then **DPO=**$P_1, P_{11}$.

Class Imbalance

When we use learning algorithms to deal with real world problems different challenges arise, such as the class imbalance problem, that is there may be a large amount of examples for one class but only a small amount of them for the class of interest. This problem may cause poor classification results. There are three general ways to treat this problem:

- The use of more adequate performance measures when facing the class imbalance problem.
- The use of techniques that bias the algorithms in order to enhance the classification performance of the minority class.
- The use of algorithms that bias the data in order to increase the amount of data for the minority class or decrease the amount of data belonging to the majority class.

Among the last alternative we find techniques such as *SMOTE* y *Resampling*. **SMOTE** (*Synthetic Minority Oversampling Technique*) is an oversampling algorithm used to increase the amount of examples of the minority class. **Resampling** references those techniques used in probability theory and statistical inference. These methods generate new samples from the observed data. They generate new simulated samples with the same size of the original sample (Good 2005); that is, the data simulation must be based on some of the real data. Then, they produce a random subsample of a dataset using a sampling with replacement technique. In this work we combined resampling with the generation of artificial images to attack the class imbalance problem in galaxy classification.

Morphological Galaxy Classification

The proposed method consists of 6 phases: (i) image segmentation, (ii) separation of stars-galaxies, (iii) nucleus separation, (iv) feature extraction, (v) artificial examples generation, and (vi) hierarchical classification.

Image Segmentation

We consider that we start from an image that contains several astronomical objects, such is the case for the plate collection at INAOE. The first step taken to classify galaxies is the stellar objects segmentation, particularly stars and galaxies. We analyzed different segmentation techniques such as border detection and thresholding. We found that a simple thresholding method worked better for this case. Thresholding methods are efficient, simple and widely used to segment and find objects of interest in an image (Stockman and Shapiro 2001).

Thresholding tries to find an intensity value, called threshold, that separates the desired classes. That is, from the image histograms, we choose a gray level that separates the corresponding values of the objects of interest from the background. The main limitation of this method is that the ideal value is not easy to find. Because of this, we used a threshold range from 120 to 150 for each of the images. These values were obtained experimentally. Finally, *very* small object are eliminated.

Stars / Galaxies Classification

An astronomical plate may contain hundreds of stellar objects. We usually find large stars populations. Because of this it is necessary to find the galaxies, by distinguishing them from star objects in the image. Then, after the image segmentation of the astronomical plate we obtained the objects of interest descriptive attributes. These attributes are the geometric moments of zero, first, and second order. These attributes describe some properties of the object such as its area and ellipticity, which are useful to distinguish galaxies from stars. We used the *Naive Bayes* and *Random Forest* classifiers implemented in Weka (Witten and Frank 2005) to perform this task and differentiate stars from galaxies.

Nucleus Separation

Once we separated stars from galaxies we focus on our objects of interest: galaxies. Our objective in this phase consists on separating the nucleus from the rest of the galaxy. According to the work described in (Lotz, Primack, and Madau 2004) we know that in the galaxies nucleus we find 20% of its flux.We used the geometric moments of zero and first order to obtain the centroid of the galaxy. Once we found the galaxy center, we calculate the galaxy Petrosian radius (Petrosyan 1982) in order to only consider the 20% of the galaxy as we show in Figure 3. This part of the galaxy is considered as its nucleus and the other 80% is considered as its body.

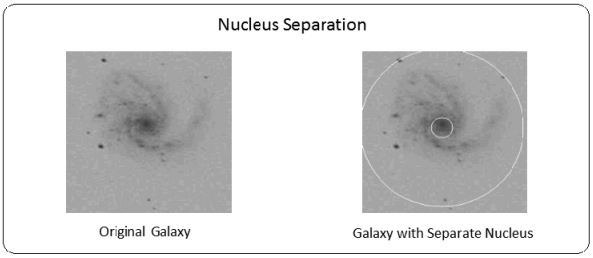

Figure 3: Nucleus Separation. Left: Segmented galaxy. Right: The two circles delimit the nucleus and body of the galaxy.

The previous step allows us to obtain the galaxy characteristics related to its brightness, such as $r_{25}r_{75}$, the rate between the nucleus and body brightness of the galaxy.

Feature Extraction

Once we segmented the image of each galaxy and separated the nucleus, the next step consists of extracting the descriptive characteristics of the galaxy. The characteristics we consider are based on those proposed by (Bazell 2000); we also used the previously calculated geometric moments as additional features.

Artificial Examples Generation

One of the most important parts of our method is the generation of artificial examples to try to compensate the class imbalance problem. In the observable universe there is a larger number of spiral galaxies than any other type.

The synthetic examples were created from the original images by applying the following geometric transformations:

- **Rotation**. It can be done in two ways. The first one with respect to the origin, that is, the position of a point is rotated around the coordinates system origin. The second way takes as reference any point (x_c, y_c) in the plane. The rotation function changes certain object characteristics such as its geometric moments of second order. In general, the rotation function is given as:

$$\begin{aligned} x^{'} &= x_c + (x - x_c)cos\theta - (y - y_c)sin\theta \\ y^{'} &= y_c + (x - x_c)sin\theta + (y - y_c)cos\theta \end{aligned} \quad (1)$$

- **Scale**. This function refers to changes in size. This transformation is also obtained in two ways. The first one is

obtained with respect to the origin. In this case, the position of the point is multiplied by a constant and it requires to specify the two scale factors S_x y S_y. In the second way we use a fixed point. This point can be the center of the object, one of its vertices, or an arbitrary point. Then, the general scaling function for an object is calculated with the following equation:

$$\begin{aligned} x^{'} &= x_c + S_x(x - x_c) \\ y^{'} &= y_c + S_y(y - y_c) \end{aligned} \quad (2)$$

The scaling function mainly affects the characteristics related to the object's area such as the geometric moments of zero order.

For the experiments, the rotation was done using angles of 45^o, 90^o and 180^o. The scaling was performed in a range from 50% to 30% depending on the size of the galaxies. We experimentally found that these parameters allow keeping most of the information of the galaxy. An example is shown in Figure 4.

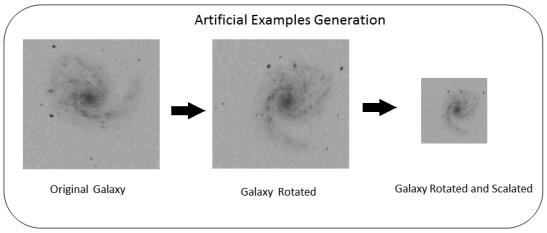

Figure 4: An example of a synthetic galaxy image. Left: original image; middle: rotation by 180^o; right: scaling by 50% of the rotated image.

Additionally to synthetic images, we also used the *resampling* method implemented in the **Weka** allowing oversampling in the percentage chosen by the user. In the experiments we compare the results using artificial examples, oversampling and both combined.

Hierarchical Classification of Galaxies

In order to perform this classification, the first step consisted of defining the hierarchy to be used for our experiments. We decided to use a hierarchy based on the Hubble's diapason diagram (the subclasses E0–E7 were not considered as the databases used in the experiments do not specify them). This hierarchy can be seen in Figure 5.

We found our best results using the multidimensional hierarchical classification method described before (Julio Hernandez 2013). The classification is based on the combination of the predicted probabilities by each classifier. As we previously mentioned, there are three ways to perform such a combination. In this work we used the product of probabilities (**PP**) alternative.

Experiments and Results

INAOE has a huge collection of astronomical plates taken by a Schmidt camera that covers a period of 50 years; analyzing these plates is part of the motivation of this work. From this

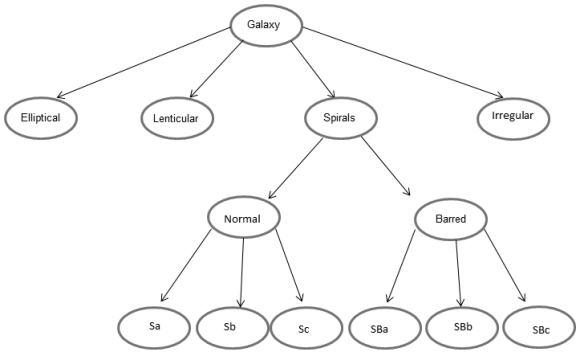

Figure 5: Galaxy taxonomy used fin the first set of the experiments.

collection we considered a subset of plates which include different types of galaxies for these experiments (digitizing the whole collection implies a great effort which is still in progress). This dataset includes a total of 152 galaxies contained in 24 astronomical plates. Where the class with the highest number of examples contains 32 galaxies and the class with less number of examples contain only 6. In order to compare our method with another using a hierarchical classification technique, we also used the database described in (Bazell 2000), which was taken from the work of (Naim et al. 1995). This second database contains a total of 834 images. We only used 800 of them because of problems due to image quality and availability.We used this database since not all work specified which were the images they used. In all the experiments we used *10-fold cross validation* in order to obtain the classifiers' accuracy. The artificial examples and oversampling we only applied to the training set, such that the testing set only includes original data.

Classification of Stars / Galaxies

As we previously mentioned, a plate may contain hundreds of stellar objects. Then we performed a first classification step in order to separate stars from galaxies. In order to achieve this classification we used the *Naive Bayes* and *Random Forest* classifiers . As the stellar objects descriptive features, we used the zero, first, and second geometric moments. The results are summarized in table 1. We observe that the Random Forest classifier provides good results, with over 90% accuracy.

Table 1: Accuracy for the star / galaxy classification using two alternative classifiers.

Classifier	**Accuracy**
Naive Bayes	79.16
Random Forest	**91.66**

Galaxy Classification

In this section we present the results obtained with the hierarchical scheme for galaxy classification of the INAOE

plates data set, and compare it with a flat classification. For each class, the number of artificial examples was proportional with the objective to have approximately the same number of examples in each class. Table 2 shows the number of classes and the types of galaxies used for the classification experiments.

Table 2: Types of galaxies used for the flat classification.

Classes	Types of galaxies
4	E, S, S0, I
9	E, Sa, Sb, Sc, SBa, SBb, SBc, S0, I

Flat Classification

The objective of this experiment is to evaluate the impact of oversampling and/or generation of synthetic images, and also to use it as a baseline for the hierarchical classifier. The tests were performed with 4 and 9 classes of galaxies. Flat classification results for four types of galaxies can be seen in table 3. In table 4 we show the flat classification results for nine types of galaxies.

Table 3: Accuracy for the flat classification of four types of galaxies.

	Naive Bayes (%)	**Random Forest** (%)
Data	25.5814	58.1395
Data and Artificial Examples	33.65	58.65
Data and Resampling (100%)	40.45	60.55
Data and Resampling (300%)	36.38	58.19
Data and Resampling (500%)	38.88	64.03
Data + Artificial Examples + Resampling (100%)	34.99	57.99
Data + Artificial Examples + Resampling (300%)	34.77	61.87
Datas + Artificial Examples + Resampling (500%)	33.60	**64.17**

Table 4: Accuracy for the flat classification of nine types of galaxies.

	Naive Bayes (%)	**Random Forest** (%)
Data	22.09	23.25
Data and Artificial Examples	26.92	30.76
Data and Resampling (100%)	29.02	27.91
Data and Resampling (300%)	27.63	31.25
Data and Resampling (500%)	26.52	29.99
Data + Artificial Examples + Resampling (100%)	25.04	38.42
Data + Artificial Examples + Resampling (300%)	31.35	33.16
Data + Artificial Examples + Resampling (500%)	34.21	**39.44**

We observe that the use of artificial examples and oversampling have a significant improvement in the performance of both classifiers, and the best results are obtained by combining both techniques with the Random Forest classifier; but the performance decreases as the number of classes is increased from 4 to 9. According with the results, the oversampling technics improve the precision percentages. However, a high percentage of oversampling doesn't mean that the precision percentages will be better since, the copies created with the resampling technic can be taken from the less representative examples, that is, examples with attributes that do not help to distinguish between classes, producing noise in the dataset.

Hierarchical Classification

For the hierarchical galaxy classification we used the multi-dimensional hierarchical classification algorithm (Julio Hernandez 2013) and the hierarchy shown in Figure 5 with 9 classes. We summarize the results in table 5.

Table 5: Accuracy for the hierarchical classification of nine types of galaxies

	Naive Bayes (%)	**Random Forest** (%)
Data	28.57	42.85
Data + Artificial	21.42	28
Data + Resampling (500%)	25	42.86
Data + Artificial + Resampling(500%)	21.43	**53.57**

We again observe that the use of artificial examples and oversampling provide a significant improvement. If we compare with the flat classification with 9 classes there is an important improvement by using a hierarchical classifier, with and without artificial examples and oversampling; the best result outperforms the best one for the flat classifier by 14 points.

Comparison with Other Approach

These experiments were performed with the second database (Bazell 2000). We extracted the features described before to these images. We also used the oversampling techniques and the creation of artificial examples. In this case we only considered six types of galaxies (E, S0, Sa, Sb, Sc, and Sd/Sm/Irr); as (Bazell 2000) used these classes of galaxies. The hierarchy used is shown in Figure 6 and the obtained results are shown in table 6.

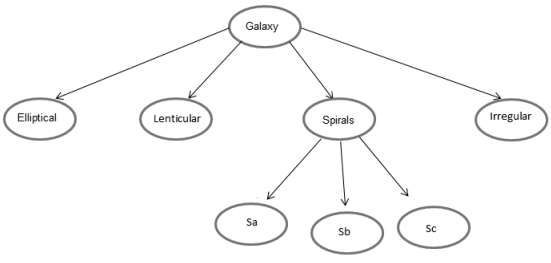

Figure 6: Hierarchy used for the experiments performed with the database of Bazell and Aha.

Table 6: Comparison of our hierarchical method with the method of (Bazell 2000).

	6 clases
Bazell and Aha	45.95
Proposed method	46.42

As the hierarchy is reduced in this case there is not so much difference between a hierarchical and flat classification. We also applied a statistical significance *T* test with confidence of 95%. In this case our method showed to be statistically superior to the method proposed by (Bazell and Aha 2001).

Conclusions and Future Work

In this work we proposed a methodology for morphological galaxy classification that includes two main contributions. The first one is the generation of artificial images of galaxies through geometric transformations to diminish the class imbalance problem. The second one is the use of a hierarchical classifier to improve the accuracy as the number of classes considered increases.

As the first step to perform galaxy classification we separated stars from galaxies. Given that the obtained results were good (more than 90% of accuracy) we conclude that the geometric moments are good features to distinguish between stars and galaxies.

From this work we conclude that the hierarchical classification method enhances the performance of flat classification methods. This enhancement is better appreciated as the number of classes increases. According to the obtained results, the flat classification accuracy results increased when we used any of the two techniques to deal with the class imbalance problem. In the case of the hierarchical classification we only obtained increased accuracy results when we applied both techniques. The results show that the use of a hierarchical classification method is a promising alternative for the galaxy classification problem.

As future work we propose:

1. Incorporate a higher number of plates to our database. In this way our database will contain a higher number of galaxies examples and the class imbalance problem will affect at a lower scale.
2. We also want to analyze and use isophotes (lines with the same brightness of an object that outline the different grey levels of the image) of the galaxies and include them in the list of descriptive features to be used for the classification task.
3. Include those galaxies types that are found in the transition between one type to the other. This could help to have more delimited types of galaxies.
4. Combine the galaxies spectral and morphological characteristics.

References

Bazell, D., and Aha, D. W. 2001. Ensembles of classifiers for morphological galaxy classification. *The Astrophysical Journal* 548(1):219.

Bazell, D. 2000. Feature relevance in morphological galaxy classification. In *Mon.Not.R. Astron. Soc.*, 519–528.

De la Calleja, J.; Huerta, G.; Fuentes, O.; Benitez, A.; Domínguez, E. L.; and Medina, M. A. 2010. The imbalanced problem in morphological galaxy classification. In *Proceedings of the 15th Iberoamerican congress conference on Progress in pattern recognition, image analysis, computer vision, and applications*, CIARP'10, 533–540. Berlin, Heidelberg: Springer-Verlag.

Freitas, A., and de Carvalho, A. C. 2007. *A Tutorial on Hierarchical Classification with Applications in Bioinformatics.*, volume Research and Trends in Data Mining Technologies and Applications. Idea Group. chapter VII, 175–208.

Good, P. I. 2005. *Resampling Methods: A Practical Guide to Data Analysis*. Birkhauser.

Julio Hernandez, L. Enrique Sucar, E. F. M. 2013. Multidimensional hierarchical classification. In *Proc. of the 26th FLAIRS, St. Pete Beach*. Florida AI Research Society.

Karttunen, H.; Kröger, P.; Oja, H.; Poutanen, M.; and Donner, K. 2007. *Fundamental Astronomy*. Springer-Verlag Berlin Heidelberg.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, 170–178. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Lahav, O. 1996. Artificial neural networks as a tool for galaxy classification. Technical Report astro-ph/9612096.

Lotz, J. M.; Primack, J.; and Madau, P. 2004. A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal* 128(1):163.

Naim, A.; Lahav, O.; Buta, R. J.; Jr., H. G. C.; de Vaucouleurs, G.; Dressler, A.; Huchra, J. P.; van den Bergh, S.; Raychaudhury, S.; Jr., L. S.; and Storrie-Lombardi, M. C. 1995. A comparative study of morphological classifications of apm galaxies.

Petrosyan, A. R. 1982. On the connection between seyfert galaxies and neighboring objects. *Astrophysics* 18:312–321.

Stockman, G., and Shapiro, L. G. 2001. *Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st edition.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Xiao, Z.; Dellandra, E.; Dou, W.; and Chen, L. 2007. Hierarchical Classification of Emotional Speech. Technical Report RR-LIRIS-2007-006, LIRIS UMR 5205 CNRS/INSA de Lyon/Universit Claude Bernard Lyon 1/Universit Lumire Lyon 2/cole Centrale de Lyon.