Content-Based Retrieval of Astronomical Images

Jorge de la Calleja, Olac Fuentes and Aurelio López-López
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro # 1
Santa María Tonantzintla, Puebla, 72840, México

Abstract

We describe an approach to perform content-based retrieval on a database of astronomical images. The method first employs image processing to normalize the images, eliminating the effects of orientation and scale, then it performs principal component analysis to reduce the dimensionality of the data, and finally, retrieval is done using the nearest neighbors algorithm. The approach has been tested on a collection of 309 images of galaxies of different types, with varying intensities, positions and orientations, yielding very good results.

Key Words:
image analysis, principal component analysis, information retrieval

1. Introduction

With numerous digital sky surveys across a range of wavelengths, astronomy has become an immensely data-rich field. For example, the Sloan Digital Sky Survey will produce more than 50,000,000 images of galaxies in the near future. This overload creates an astringent need for techniques to automate the search in such a huge volumes of data. Another problem is that astronomical images are commonly noisy with objects of diffuse nature. It has been realized that techniques specific to face this problem are needed [3].

In order to handle the enormous volume of data, not only of the big number of images but also the size of each image file, it is necessary to summarize the information. From this perspective, we propose an approach to perform image retrieval that initially generates a representation that is independent of scale and orientation, then generates a more compact representation, amenable to exhaustive search, using principal component analysis. These processes set the stage to receive a query image and compare it against those in the collection, using Euclidean distance in the eigenspace generated, to find the images that are most similar to it.

The paper is organized as follows: Section 2 describes the general architecture of the system, including the vision, data compression and image retrieval modules. In Section 3, we present experimental results, describing the application of the system to a data base containing 309 astronomical images, depicting galaxies of different types with varying brightnesses, positions and orientations. Finally, Section 4 presents conclusions and suggest some directions for future research.

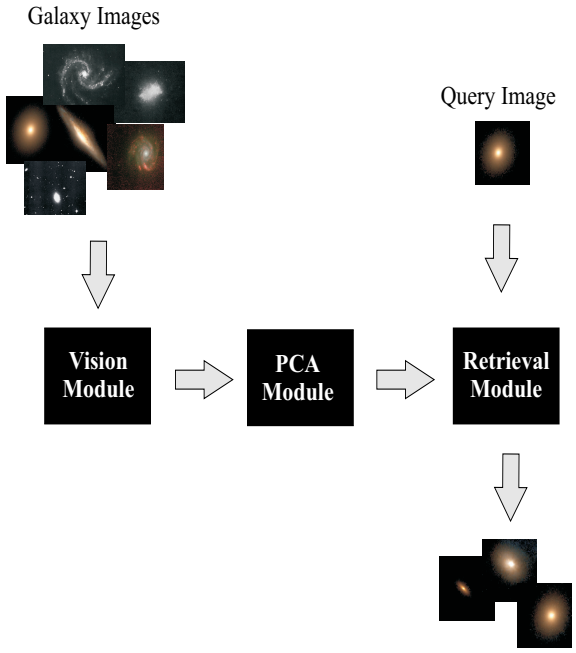

Figure 1. Modules of the Retrieval System

2. Description of the Approach

The method that we developed for image retrieval of galaxies is divided in three modules (see Figure 1): the Vision module, the Principal Component Analysis (PCA) module, and the Retrieval module (engine). The method works as follows: it takes as input the galaxy images, then the Vision module rotates, centers, and crops them; the PCA module finds the eigenvectors, and the projection of the images onto the principal components will be the input parameters for the Retrieval module. At the end, the user can supply a query image and the Retrieval module, after processing in the same way the image, compares it against those in the collection, producing as a response the images found that are similar to the query image. The next three subsections describe in detail each module.

2.1 Vision Module

The images of galaxies are usually of different sizes, colors and formats, and most of the time, the galaxy contained in the image is out of the center. So, the aim of this module is to create images that are invariant to color, position, orientation and size. So, the vision module rotates, centers, and crops the galaxy images, all in a fully automatic manner. First, we find the galaxy contained in the image applying a threshold, that is, we generate a binary image B, such that

$$B(i,j) = \begin{cases} 1 & \text{if } I(i,j) > \text{threshold} \\ 0 & \text{otherwise} \end{cases}$$

Then we obtain $\bar{i}$ and $\bar{j}$, the center row and column of the galaxy in the image, given by

$$\bar{i} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} iB(i,j)$$

$$\bar{j} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} jB(i,j)$$

where m and n are the number of rows and columns, respectively. Then we obtain the covariance matrix of the points in the galaxy image

$$C = \sum_{i=1}^{m} \sum_{j=1}^{n} B(i,j)[i-\bar{i}, j-\bar{j}]^T [i-\bar{i}, j-\bar{j}]$$

The galaxy's main axis is given by the first eigenvector (the eigenvector with the largest corresponding eigenvalue) of C, the covariance matrix. We then rotate the image so that the main axis is horizontal (see Fig 2). The angle is given by

$$\theta = arctan(PC1[1]/PC1[2])$$

where PC1 is the first principal component. Then we use an image warping technique to rotate the image.

After that, we crop the image, eliminating the columns that contain only background (black) pixels. Finally, we stretch and standardize the images to a size of 128x128 pixels. Figure 2.1 shows examples of the image processing stage for three galaxies that representing the three main types included in the standard Hubble classification scheme, an elliptical galaxy, a spiral galaxy, and an irregular galaxy.

2.2 Principal Component Analysis Module

The basic idea in PCA is to find the components (the eigenvectors) of the covariance matrix of the set of images, so that they explain the maximum amount of variance possible by n linearly-transformed components. These eigenvectors can be thought of as a set of features which together characterize the variation among the images [6]. This module

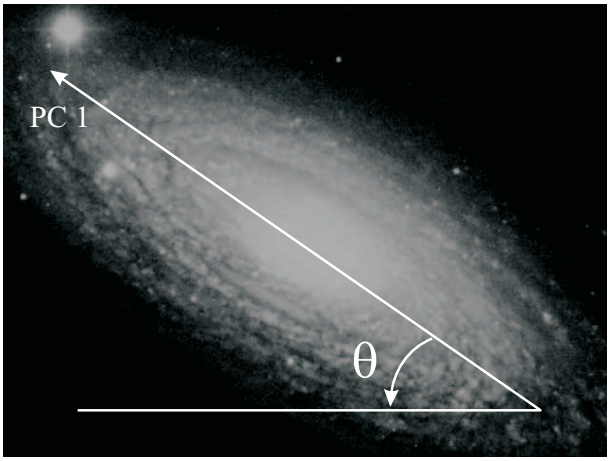

Figure 2. Rotation of the galaxy using the first principal component (PC1).

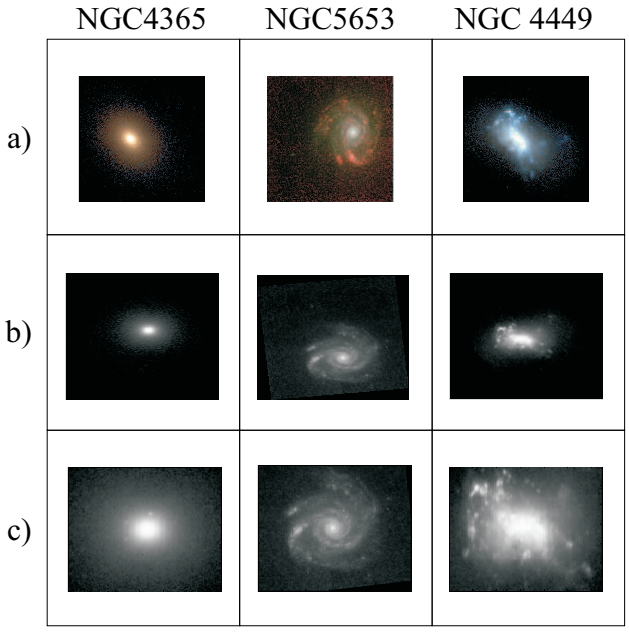

Figure 3. Examples of Galaxies, a) Original images, b) Rotated images, and c) Cropped and centered images.

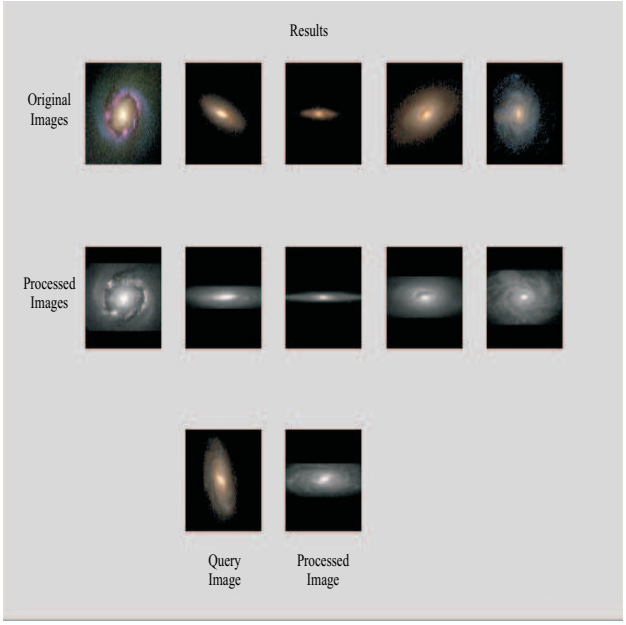

Figure 4. First Example of Galaxy Retrieval.

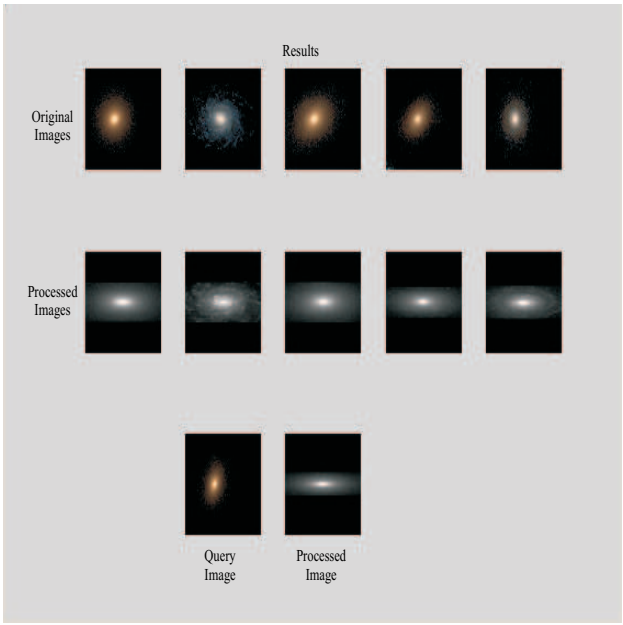

Figure 5. Second Example of Galaxy Retrieval.

takes as input the set of processed images, and finds the principal components (PCs). The projection of the images onto the PCs are taken as the representation used by the retrieval engine to compare against the query image. We have used 25 PCs because they represent about 85% of the information in the image set and yield good results both in terms of accuracy and running time.

2.3 Retrieval Module

The Retrieval Engine takes as input the collection of images processed as detailed in 2.1 and 2.2 and a query image provided by the user. The query image is processed in the same way as the original images, that is, it is centered, rotated, and cropped, and then its projection onto the PCs is obtained. Then, once having the query image represented in the same space as those of the collection, its nearest neighbors are found, using the standard Euclidean distance. The images closest to the query are taken and displayed as a result. In the current implementation, we perform brute-force search to find these nearest neighbors, which can be done quickly given the small size of the database. However, when we have to deal with larger databases, a more efficient search method such as KD trees will have to be applied.

3. Experimental Results

We tested the system using a data set that consisted of 309 images of galaxies. It was taken from the NGC catalog on the web page of the Astronomical Society of the Pacific [4]. We processed the images as detailed in Section 2. Figure 3 shows examples of images from the original data set and the resulting images output by the vision module.

In order to assess the effectiveness of the approach, we used leave-one-out cross validation, testing the output of the system with one image, and training with the rest, and repeating the process 309 times, until all images have been used once as test image. Using this approach, the system output as best match a galaxy belonging to the same class as the query image 89% of the time, considering the three main galaxy types defined by the Hubble system: spiral, elliptical and irregular. This error rate is smaller to those reported in the literature using automated means for image-based galaxy classification ([5, 1]. The best results were obtained when using an elliptical galaxy as query image, very likely because they present the most regular structure, while the worst results were obtained for irregular galaxies, which have very little discernible structure.

In Figures 4, 5, and 6, we show three examples of retrieval. In these figures, the five images closest to the query are given as a results of retrieval. The first row displays the original images, the second row includes the images after processing, and the third row includes the query image before and after processing.

4. Conclusions

We have presented a system that performs content-based retrieval of astronomical images. The system executes the following steps to perform image retrieval:

1. Use computer vision techniques to find the location, orientation and size of the galaxy in the image.

2. Rotate, crop and resize the images so that in all the

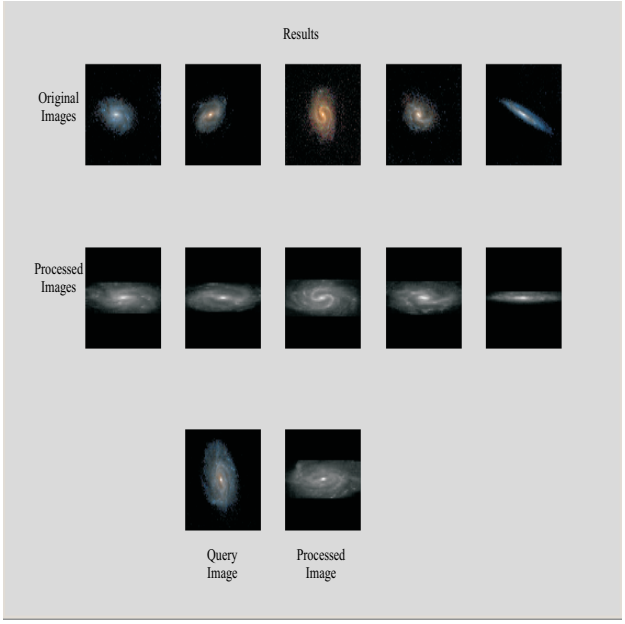

Figure 6. Third Example of Galaxy Retrieval.

images are the same size, the galaxy is at the center of the image, has horizontal orientation and covers the whole image.

3. Find the eigenvectors of the images and project the images into the eigenspace formed by the first 25 eigenvectors.

4. Given a query image, process it as in steps 1 and 2, project it onto the eigenspace obtained in 3 and retrieve the n images with the smallest Euclidean distance to it in the eigenspace.

Quantitative results show that almost 90% of the time the image deemed by the system as most similar to que query belongs to the same class, and qualitative results show that the set of images retrieved by the system are visually similar to the query image.

Some directions of future work include:

- Extending the experiments to a larger database of galactic images
- Building classifiers for other types of astronomical objects, such as nebulas and clusters.
- Extending the system to deal with wide-field images, containing multiple objects. This will be done by means of a preprocessing stage to segment the objects in the images, and then processing them individually.

5. Acknowledgments

We would like to thank CONACyT for partially supporting this work under grant J31877-A and fellowship 166596 for the first author.

References

[1] A. Adams and A. Woolley, Hubble Classification of Galaxies Using Neural Networks, Vistas in Astronomy 38, 1994.

[2] J.L. Bentley, Multidimensional Binary Search Trees Used for Associative Searching. Communications of the ACM, 18(9), September 1975.

[3] Csillaghy, A., Hinterberger, H., Benz, A.O.: Content-based Image Retrieval in Astronomy. In: Kluwer, Netherlands (2000) 1–16

[4] NGC Images on the Web page of the Astronomical Society of the Pacific. http://www.apsky.org/ngc/ngc.html

[5] M. C. Storrie-Lombardi, O. Lahav, L. Sodre and L. J. Storrie-Lombardi, Morphological Classification of Galaxies by Artificial Neural Networks, Monthly Notices of the Royal Astronomical Society 259, 1992.

[6] M. Turk and A. Pentland, Face Recognition Using Eigenfaces. Proceedings of IEEE CVPR, 1991.