

NCRA-TIFR PROJECT PROPOSAL

SUPERVISED LEARNING & ANALYTICAL MODELING FOR AUTOMATED MORPHOLOGICAL CLASSIFICATION OF RADIO SOURCES



Archit Sakhadeo
Rathin Desai
Shadab Shaikh
Shubhankar Deshpande

Mentor
Dr. Yogesh WADADEKAR
Dr. C. H. Ishwara CHANDRA

1 Introduction

1.1 Morphological Classes of Radio Galaxies

Radio galaxies with active nuclei can be distinguished based on their radio luminosity or brightness of their radio emissions in relation to their hosting environment. Some of the basic morphological classifications include point sources, extended sources i.e. sources with extended contours, double radio sources, jets, and lobes.

1.2 Problems faced with current classification

Currently Radio astronomers manually classify galaxies based on visual inspection of the images which is a slow procedure, and increases the time to production of scientific results. Further, it introduces uncertainties in the classification procedure, both of which are problems which can potentially be mitigated by using an automated approach.

Contemporary algorithms classify radio sources into at most three different classes. Our aim is to build a robust model capable of handling more than 2 classes.

2 Objective

- Potentially discovering rare forms of radio sources by classification in different classes.
- Reduction in time to generate scientific results by radio astronomers.
- Deeper insight into topological representation of radio data during classification.

3 Approach

3.1 Source Modelling

The first step would be source extraction using the standard technique of gaussian modelling. We propose to do this using the robust PyBDSM pipeline used for fitting gaussian distributions to radio sources. The software contains a plethora of features, from which we would be using a small subset. This would mainly include:

1. Source extraction using gaussian modelling of radio data.
2. Generation of a catalog file containing details of radio sources (RA, DEC, Size of Gaussian (min, max), etc.)

3.2 Cutout Generation

The second step would be to convert the RA(Right Ascension) and DEC (Declination) values generated from the catalog, to their corresponding pixel values in the original image. Based on these pixel values we generate 10*10 px cutouts using as reference the co-ordinates of the center of the radio source. This involves a multistep procedure briefly including:

1. Reading the FITS image in the form of a matrix
2. Parsing through the generated catalog file and extracting data for each radio source such as RA, DEC, etc.
3. Converting the RA and DEC values from WCS (World Coordinate System) to pixel values.
4. Processing pixel values to account for difference in addressing between FORTRAN and C family of languages.
5. Slicing the image matrix assuming the reference pixel co-ordinates as the center of the source.

Prototype code for section 3.1 and section 3.2 has been written mainly for testing purposes. We used a sample image from the TGSS survey which was then processed using the first two steps of our pipeline to generate 470 cutout images. More details can be found at: <https://github.com/NCRA-TIFR/radiogen>.

3.3 Data Preprocessing

Real world data is incomplete, inconsistent and noisy. Techniques like data cleaning, integration, transformation, reduction, discretization are required to ensure data uniformity.

Standard techniques include:

1. Mean Subtraction

The mean is subtracted across every feature in the data cloud. The geometric interpretation of this operation is the centering of the data cloud around the origin along every dimension.

2. Normalization

The data dimensions to span the same range. This results in higher accuracy during data processing, as inter-dimensional variance is accounted for. There are two common ways of achieving this normalization.

- (a) Divide each dimension by its standard deviation, once it has been zero-centered.
- (b) Normalize each dimension to result in a range of values [-1,1].

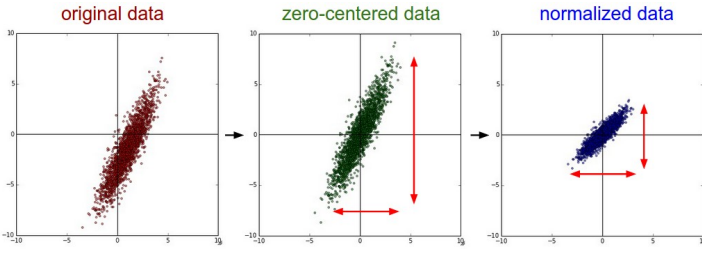


Figure 1: Mean Subtraction and Normalization

3. Dimensionality Reduction using Principal Component Analysis

The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. By finding the eigen vector with the highest eigen value we select the Principal component axis with the highest variance and the minimum reconstruction error. Dimension reduction occurs by eliminating axes with eigen vectors exhibiting the least variance. This results in a data representation exhibiting a higher degree of variance, leading to faster convergence and greater accuracy during classification.

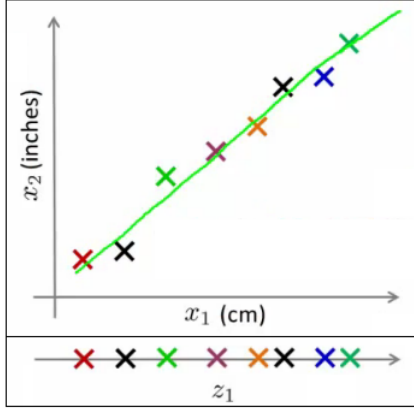


Figure 2: Reduction from 2D to 1D using PCA

3.4 Analytical Approach

The analytical approach to solving the problem of galaxy morphology classification would primarily contain two steps:

1. Feature extraction using sophisticated techniques prevalent in image processing.
2. Building a statistical model of the extracted features
3. Cross comparison of the designed model with principal indicators of dominant classes

A variety of techniques have been used to tackle the related problem of visual recognition in Computer Vision. The standard algorithm for feature extraction is the Scale Invariant Feature Transform (SIFT).

3.4.1 Scale Invariant Feature Transform

The SIFT algorithm extracts keypoints from a set of reference images and stores them in a database. The feature vectors of the extracted images are compared with the reference vectors on the basis of a defined norm (usually Euclidean distance). The probability that of the subset of features indicating the presence of an object is computed. This gives the accuracy of the fit. When multiple features pass these tests, the class can be correctly identified with a high confidence.

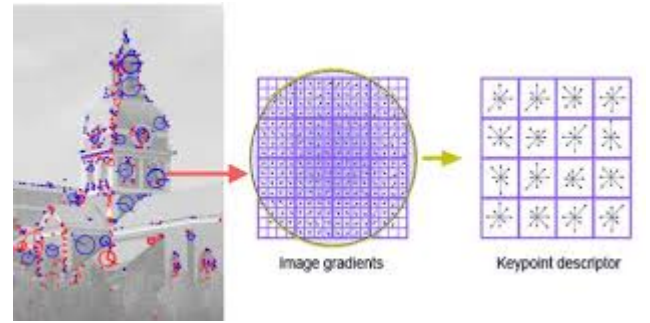


Figure 3: Feature Extraction using SIFT

3.5 Empirical Approach

An empirical approach is a way of gaining knowledge by means of direct and indirect observation or experience. In the context of this project, we plan to employ Artificial Neural Networks(ANNs), which are a computational model based on a large collection of artificial neurons, that learn hierarchical representations of the data. Each neural unit is connected with many other units which computes to essentially create a non-linear stacked function approximator. Neural networks typically consist of multiple layers in which a signal traverses from the input layer to the output layer.

A convolutional neural network (CNN), is a type of ANN, is a state of the art model that currently provide the best solutions to many problems in computer vision such as image segmentation, recognition and classification[1].

Convolutional neural networks were designed to use minimal amounts of preprocessing. A typical architecture of a CNN consists of convolutional layers, activation layers and pooling layers each of which is briefly explained.

- Convolution layer: Performs the convolution operation with a static filter on input from previous layers.

- Activation layer: Defines the output of a unit given an input or set of inputs by applying an activation function.
- Pooling layer: Progressively reduces the spatial size of the representation, which results in a decreased number of parameters and thus reduction in network computation.

The initial layers of the CNNs learn simple features in the input data such as straight edges, simple colors, and curves. The deeper layers learn the higher level representation of image and search for complex shapes and structures. Mathematically, deeper layers can be thought of as compositions of previous layers. [1][2][3].

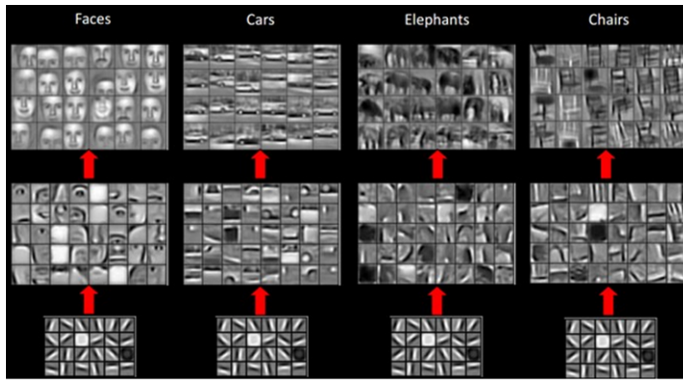


Figure 4: CNN learning simple features on initial layers for different classes

Variants of CNNs have achieved great successes for morphological galaxy classification and prediction on SDSS[4]. Sander Dieleman et al. won the Galaxy challenge, an international competition on Kaggle, to build the best model for morphology classification based on annotated images from the Galaxy Zoo project.

4 Timeline

April - May	• Literature survey
Mid-August to September	• Hypothesis testing
September to October	• Prototype implementation
October to November	• System refinement
November to December	• Publishing results

5 Conclusion

We would like to thank Dr.Yogesh Wadadekar ¹, Dr.C. H. Ishwara Chandra ¹ for their supportive presence during

the process of brainstorming potential research ideas. Their constant guidance has been an invaluable source of inspiration for us, and we are eager to continue working with them.

References

- [1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*.
- [2] Matthew D. Zeiler, Rob Fergus. *Visualizing and Understanding Convolutional Networks*. arXiv print : 1311.2901v3, 2013.
- [3] Andrej Karparthy, Fei-Fei Li, Justin Johnson, Serena Yeung. <http://cs231n.github.io/convolutional-networks/>
- [4] Sander Dieleman, Kyle W. Willett and Joni Dambre *Rotation-invariant convolutional neural networks for galaxy morphology prediction*. arXiv print : 1503.0707v1, 2015.

¹ National Center for Radio Astrophysics - Tata Institute of Fundamental Research