

GEOMETRIC METHODS IN ASTRONOMICAL IMAGE
PROCESSING

M.Sc THESIS

SUBMITTED AT THE UNIVERSITY OF LUENEBURG
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

Haider Ali
April, 21, 2006

**Statement concerning the final thesis
for the master's programme in Software Technology (M.Sc.)**

Last Name: First Name:

Student ID Nr.:

I / we affirm

1. that I / we have produced this thesis on our own and without using any other means than those explicitly identified,
2. that all ideas in this thesis which have been taken verbatim or in other phrasing from publications or other sources have been correctly cited.

Lüneburg, (date)

.....
Signature/s

Preface

”Astronomy is perhaps the science whose discoveries owe least to chance, in which human understanding appears in its whole magnitude, and through which man can best learn how small he is.”

G. C. Lichtenberg (1742 - 1799) German physicist

”Let no one expect anything of certainty from astronomy, lest if anyone take as true that which has been constructed for another use, he go away . . . a bigger fool than when he came to it.”

Nicolas Copernicus (1473 - 1543) Polish astronomer

This thesis describes a new system for ”Finding Satellite Tracks” and ”Measurement of Galaxy Ellipticity” based on the modern geometric and statistical approaches. There is an increasing need of using methods with solid mathematical and statistical foundation in astronomical image processing. Where the computational methods are serving in all disciplines of science, they are becoming popular in the field of astronomy as well. Currently different computational systems are required to be numerically optimized before to get applied on astronomical images. So at present there is no single system which solves the problems of astronomers using computational methods based on modern approaches. The system ”Finding Satellite Tracks” is based on geometric matching method ”*Recognition by Adaptive Subdivision of Transformation Space (RAST)*” and ”Measurement of Galaxy Ellipticity” based on the statistical method ”*Maximum Likelihood Estimation*”.

Acknowledgments

This report is a result of my research activities at the Research Lab of Image Understanding and Pattern Recognition (IUPR) at German Research Center for Artificial Intelligence and Technical University of Kaiserslautern, Germany. I would like to thank my supervisor Prof Dr T. Breuel with my heartiest gratitude for giving me the opportunity to work with him, for his great interest and encouragement in this work. I would also like to thank my principal advisor Prof. Dr J. Jacobs for his interest in my research. I appreciate his valuable and constructive suggestions for this work. My special thanks to Dr C. Lampert and Dr T. Erben for their kind support and guidance during all my research activities at the Institute of Image Understanding and Pattern Recognition.

I have many thanks to Faisal Shafait and Joost Beusekom throughout fruitful discussions as they helped me in one way or other way around while implementing my ideas.

Finally my heartiest gratitude to my Mother. Without her love, support and understanding, I would not have succeeded to complete this work.

in the memory of my father

Contents

Preface	iii
Acknowledgments	iv
1 Introduction	1
1.1 Motivation	1
1.2 Problem Formulation	2
1.2.1 Finding Satellite Tracks	2
1.2.2 Measurement of Galaxy Ellipticity	2
1.3 Related Work	3
1.3.1 Satellite Tracks Removal	3
1.3.2 Measurement of Galaxy Ellipticity	3
1.4 Thesis Organization	4
2 Finding Satellite Tracks	5
2.1 Image Pre-Processing	5
2.1.1 Conversion	5
2.1.2 Morphological Operation	6
2.2 Geometric Matching	8
2.2.1 General Description	8
2.2.2 Data Preparation	10
2.2.3 Recognition by Adaptive Subdivision of Transformation Space (RAST) Algorithm	10
2.2.4 Geometric Transformation	10

2.2.5	Parametrization	11
2.2.6	The Algorithm	13
2.2.7	Output	14
2.3	Image Reconstruction	14
2.3.1	General Description	14
2.3.2	Painting Image	14
2.4	Summary	17
3	Measurement of Galaxy Ellipticity	18
3.1	Image Pre-Processing	18
3.1.1	Conversion	18
3.2	Detection of Objects	18
3.2.1	Connected Components	18
3.3	Measurement of Galaxies	20
3.3.1	Gaussian Grayscale Distribution	20
3.3.2	Maximum Likelihood Estimation	20
3.3.3	Image Reconstruction	22
3.4	Summary	25
4	Evaluation	26
4.1	Removal of Satellite Tracks	26
4.1.1	Results	26
4.2	Measurement of Galaxy ellipticity	30
4.2.1	Results	30
5	Conclusion and Future Work	35
	Bibliography	37

List of Tables

4.1	Different types of errors made by satellite tracks detection application. Each column represents the different type of error. The column labels are: total true positive (T_p), total true negative (T_n), total false positive (F_p), total false negative (F_n)	27
4.2	Different types of results made by RAST algorithm while finding satellite tracks in the images. Each column represents different type of error. The column labels are: true positive (T_p), true negative (T_n), false positive (F_p), false negative (F_n)	29

List of Figures

2.1	Example of binarized image	7
2.2	Example of opening effect	9
2.3	RAST line fitting	11
2.4	Representation of r and θ in xy -plane	12
2.5	Example: Difference images of satellite tracks removal	15
2.6	Example: Difference images of satellite tracks removal	16
3.1	Example of bounding boxes	19
3.2	An example of Gaussian grayscale distribution	23
3.3	Example: Original and reconstructed galaxies	24
3.4	Example: Original and reconstructed galaxies	24
4.1	Difference images containing light spikes	28
4.2	Example of galaxies reconstruction	31
4.3	Example of galaxies reconstruction - continued	32
4.4	Example of error shapes reconstruction	33
4.5	Example of error shapes reconstruction - continued	34

Chapter 1

Introduction

1.1 Motivation

Astronomical images play an important role in man's effort to understand the universe. These images are taken by space observatories spread across the globe. Acute weather conditions are required for image capturing.

It takes several minutes to take a single astronomical image. Therefore it is important to extract maximum information from the astronomical images. These images contain important information about stars and galaxies and irrelevant information (like satellite tracks) which appear quite often. These irrelevant information create problems for astronomers in analyzing the images. Since data collection is time consuming, the images containing additional objects can not be simply discarded. Instead, irrelevant objects have to be removed from these images [6].

Currently astronomers remove these satellite tracks manually from images according to [5]. We replace this traditional approach by using a computational system. This system is based on a *Geometric Matching* technique called "*Recognition by Adaptive Subdivision of Transformation Space (RAST)*". A working system will help in saving a lot of human effort to remove these tracks manually.

The most important objects for astronomers in astronomical images are galaxies. According to [8] galaxies occur in three different shapes: elliptical, spirals and irregulars. The most common shape is elliptic. The scope of this work is limited to the

measurement of elliptic shape galaxies. "*Maximum Likelihood Estimation*" is used to measure elliptical galaxies by estimating parameters. From these parameters a shape of ellipse could be measured by estimating the *major* and *minor* axes in the *xy*-plane.

A synthetic image based on estimated parameters containing elliptical galaxies is generated for visualization and further analysis by astronomers.

1.2 Problem Formulation

The objective of this thesis work is based on two main tasks: to find satellite tracks in the images taken by astronomers and to measure elliptic shape galaxies.

1.2.1 Finding Satellite Tracks

The images taken from the sky are contaminated by a lot of noise (Cosmic rays, CCD defects, ghost images and satellite tracks) which disturbs the detection or identification of important objects in the images like stars and galaxies. In these images satellite tracks can be characterized as straight lines. The major aim of this part of work is to develop a system to find and remove satellite tracks in astronomical images. In this system the RAST algorithm is applied which is based on Geometric Matching techniques.

1.2.2 Measurement of Galaxy Ellipticity

This part of thesis work is based on machine learning techniques in order to determine the shapes of galaxies. The aim of this work is to calculate size, position and rotation of elliptical galaxies. In this work the popular statistical method Maximum Likelihood Estimation is applied to estimate mean, covariance and scale factor of a *Gaussian grayscale distribution*.

1.3 Related Work

1.3.1 Satellite Tracks Removal

There is no computational method to reliably identify and remove satellite tracks at present. Currently astronomers do this job by hand after selecting and masking the images having satellite track. After masking satellite tracks they ignore the masked area while doing further processing of the objects [5].

A general adaptive method *Cleaning Sky Survey databases using Hough Transform and Renewal String Approaches* was introduced by A.J. Storkey et al. in 2004 to remove unusual objects in the astronomical images. This method was developed to remove four types of objects - satellite or aeroplane tracks, scratches, fibers and other linear phenomena introduced to the plate, circular halos around bright stars due to internal reflections within the telescope and diffraction spikes near to the bright stars in the sky survey data using catalogue of objects as described in [13]. The system we present is different from the adaptive method, as our system uses pixel based representation of images instead of using catalogue of objects.

1.3.2 Measurement of Galaxy Ellipticity

SExtractor, an astronomical source extractor system, is in use of astronomers for building catalogue of the objects from astronomical data. This system provides the feature to measure the basic shape and ellipse parameters as described in [1]. But this system is not extendable to measure different shapes of galaxies and stars e.g. when objects overlap.

A. Refregier and D. Bacon introduced a new formalism named *Shapelets* which provides a way to measure galaxy shapes. As defined in [12], Shapelets are a complete, *Orthonormal* set of 2D basis functions constructed from Laguerre or Hermite polynomials weighted by a Gaussian. This method describes shape of each object with a series of localized basic functions. For estimating the complex shapes like overlapping galaxies, it is difficult to decide whether it is one or two overlapping objects using this method. While our system is an adaptive system which can easily

describe these complex shapes. The method we present describes each object with the shape parameters instead of describing an object with a series of localized basic functions. The presented system measures only the elliptic shape parameters and objects could be identified on the basis of these parameters. The presented system is also computationally less expensive and is based on statistical methods to solve this problem. The presented system measures galaxy ellipticity and is easily extendable to measure different shapes and rotations of elliptical as well as other shapes of galaxies and stars.

1.4 Thesis Organization

This thesis is organized as follows: Chapter 2 gives a brief overview of image pre-processing and related work including data preparation for satellite tracks system. Then the use of RAST algorithm in the system is described for finding straight lines in the image. Chapter 3 explains the procedure of parameter estimation for elliptical galaxies. It covers the statistical method Maximum Likelihood Estimation for estimating the parameters. Chapter 4 presents evaluation of the tasks and Chapter 5 leads to the discussion of conclusions and future work.

Chapter 2

Finding Satellite Tracks

2.1 Image Pre-Processing

2.1.1 Conversion

The available astronomical images are generally in Flexible Image Transport System (FITS) format. This format is widely used in astronomy for convenient exchange of astronomical data. FITS images are composed of a sequence of header data units (HDUs). A header data unit contains keywords (value statements) that describe the organization of the data in HDUs and the format of the data contents. Based on the special structure of these header data units, FITS images may provide additional information like instrument used, status and history of the data etc. [9].

The following pre-processing steps are performed on the given images before applying RAST.

Binarization

Binarization is the process of converting a grayscale image to a black and white image. A grayscale image contains pixel intensity range of 0 to 255 levels. The given images are binarized and converted from FITS grayscale images to Portable Greymap (PGM) images. In this work the binarization is done by using global thresholding.

Global thresholding sets all pixels above a defined value to white and the rest of the pixels to black in the image. It is very important to decide the appropriate threshold value to binarize the image, though it is difficult to decide a global value which is suitable for all images [10]. In this case the intensity range does not vary much from image to image. After looking at grayscale values in different images on the available data set, 102 were founded as a suitable value (40 %) threshold of total intensity range 0 – 255.

`Convert`, a utility in the ImageMagick software package was used to binarize the grayscale images.

A C++ routine was written which calls the `Convert` command. It walks through the current and subdirectories and converts FITS grayscale images to binary images. An example image before and after binarization is shown in Figure 2.1.

2.1.2 Morphological Operation

After binarization the morphological operation *Opening* was applied to remove noise and CCD defects in the images.

Opening

Opening is based on the morphological operations *Erosion* and *Dilation*. Opening smooths the inside of the object contour, breaks narrow strips and eliminates thin portions of the image. It is done by first applying erosion and then dilation operations on the image.

Erosion shrinks the foreground objects in the image by a certain amount. For two sets A and B the erosion of A by B in Z^2 is defined as:

$$A \ominus B = \{z | (B)_z \subseteq A\} \quad (2.1)$$

The set A represents foreground pixels in the original image and set B is referred as a structuring element. The erosion of A by B is the image containing all the

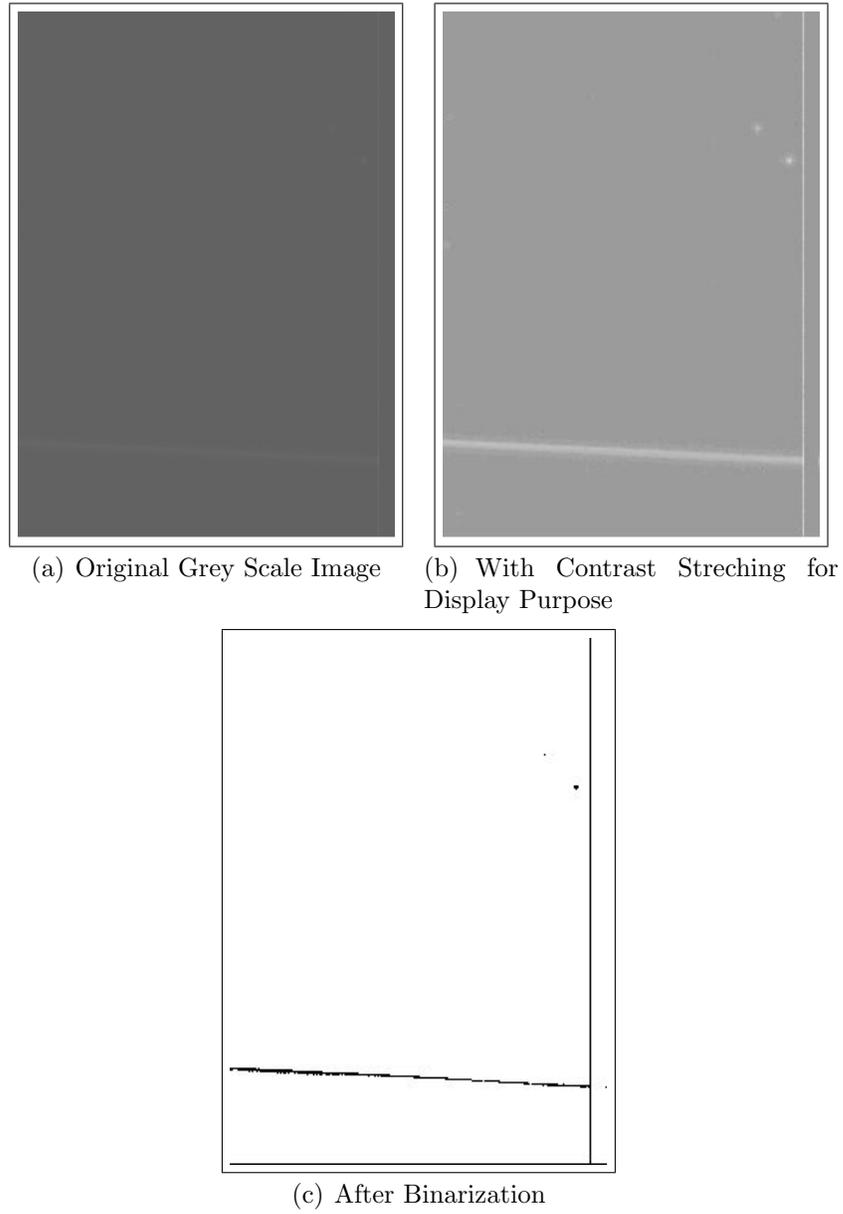


Figure 2.1: Example of binarized image (foreground objects become visible), the binarized image is inverted to white background and black foreground

foreground pixels, z , in A such that the translation $(B)_z$ of B by z is completely contained in A .

Dilation dilates foreground objects in the image. For sets A and B where B is a symmetric structuring element, dilation of set A by B in Z^2 is defined as:

$$A \oplus B = \{z | (B)_z \cap A \neq \emptyset\} \quad (2.2)$$

The dilation of A by structuring element B translated by z such that B is overlapped by at least one element in set A [10].

The amount of growing and shrinking the size of objects depends on the structuring element. CCD defects and noise have to be eliminated. These defects have a typical size of one to three pixels. So the opening is performed by using 3×3 structuring element.

`Pgmorphconv` is used for the opening operation. It is a utility in the Netpbm software package, a graphics programming library. The opening effect is demonstrated in Figure 2.2.

2.2 Geometric Matching

2.2.1 General Description

There are different algorithms to perform geometric matching. Many geometric matching problems in computer vision are based on image models. A model of an image can have geometric features like points, lines, arcs etc. Different algorithms are developed to match these geometric features in the model image [4]. As satellite tracks have geometric resemblance to straight lines, we look for line features in the image. In this work the RAST algorithm is applied for finding satellite tracks in the image.

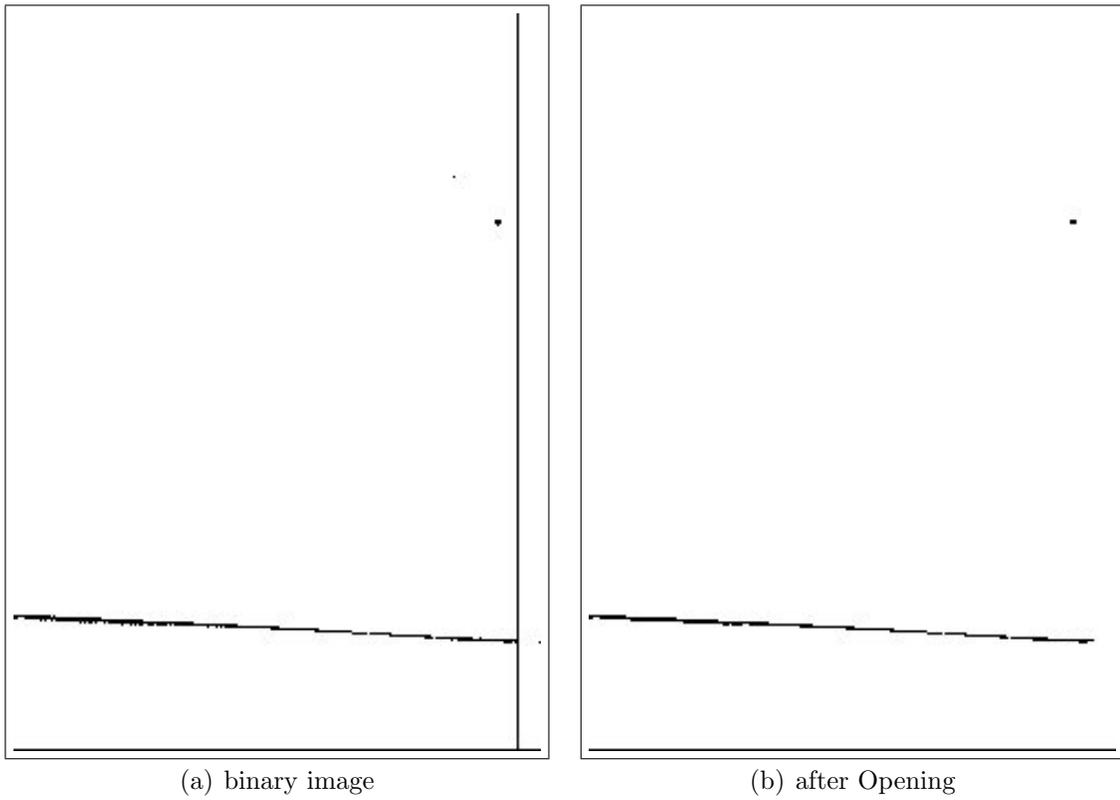


Figure 2.2: Example of opening effect (CCD defect and bad pixels are removed), Both the images are inverted to white background and black foreground

2.2.2 Data Preparation

RAST takes sample points as input features. The computation of these sample points is described in this section. In binarized images these input features consist of the pixel positions of the foreground objects.

The binarized images contains white foreground objects on a black background. A routine is written in C++, which computes the (x, y) positions of all the white pixels in the image. These pixels constitute the sample points.

Following are the steps involved in computing pixel positions in the image.

```
. Read the converted binary image
. For each pixel (x,y)
.   if pixel = white then
.   Write (x,y) pixel position into the data file
```

2.2.3 Recognition by Adaptive Subdivision of Transformation Space (RAST) Algorithm

The RAST algorithm was developed by Breuel [2]. RAST consists of a family of geometric matching algorithms, one of these is for finding lines.

2.2.4 Geometric Transformation

The RAST algorithm is applied on prepared sample data points. RAST is based on forward transformation (model to image) and an error model [3]. In this case, the algorithm takes a collection of sample data points and tries to fit an optimal line on these sample points in the image.

The algorithm is implemented using hierarchical and adaptive subdivision of the space of line parameters. RAST implements a "best first" search by a binary tree based on recursive subdivision of the parameter space [3].

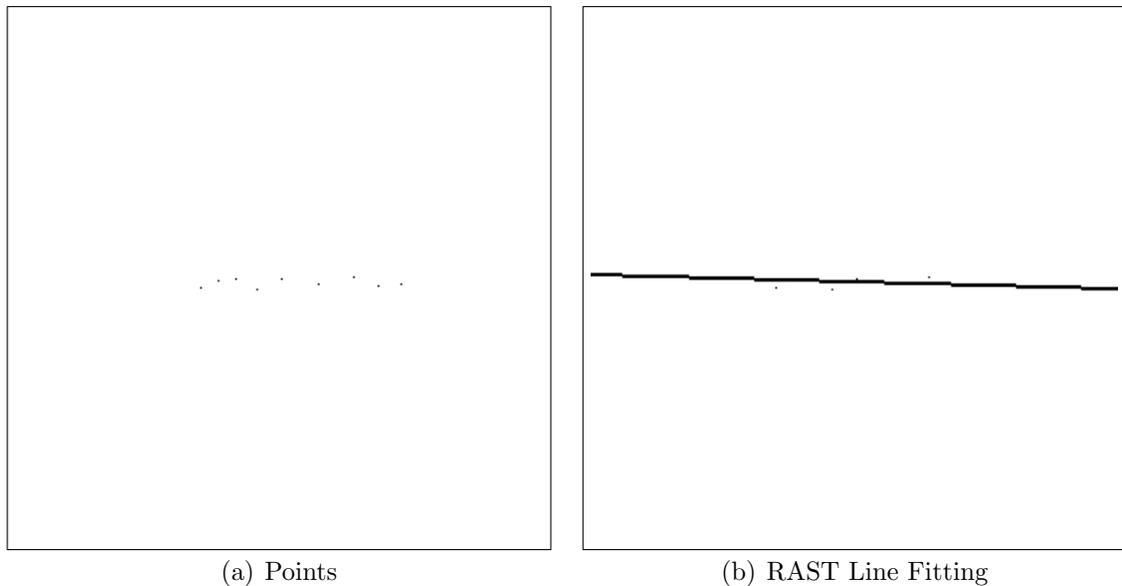


Figure 2.3: A hand made image with sample data points and reconstructed copy of the hand made image with RAST line fitting

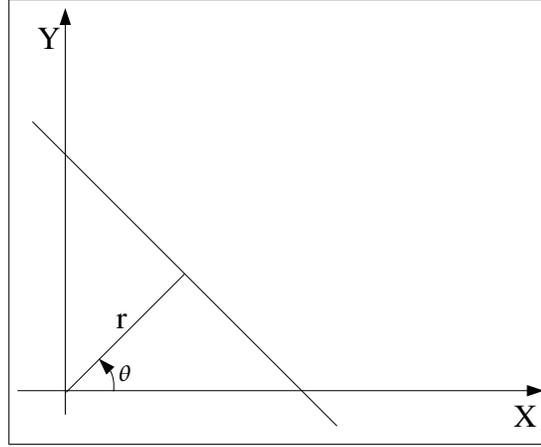
The RAST algorithm considers a set of all possible parameters as region of interest before execution. Then it recursively divides the parameter space into sub regions. For each region in the parameter space there is a set of consistent image features under bounded error. During the execution RAST eliminates the regions of disinterest. These are the regions which do not contain a solution [3].

An illustration of sample points and RAST line fitting is presented in Figure 2.3.

2.2.5 Parametrization

The RAST algorithm for finding lines based on the parametric model, described by line parameters (r, θ) . In the xy -plane, r is the distance of the line from the origin and θ is the angle between the perpendicular and the x -axis. The angle θ is always measured anticlockwise from the positive side of the x -axis. As shown in 2.3.

$$x\cos\theta + y\sin\theta = r \quad (2.3)$$

Figure 2.4: Representation of r and θ in xy -plane

An illustration of line parameters in xy -plane is presented in Figure 2.4.

There are different parameters to the algorithm like the *epsilon*, *quality*, *tolerance* and *angle tolerance*.

- The epsilon (eps) parameter defines the distance up to which a point can contribute to the line.
- The quality (minquality) parameter specifies the minimum acceptable quality of a line as given by:

$$q(\vartheta, P) = \sum_{k=1}^N \max\left(0, 1 - \frac{d_k^2}{\epsilon^2}\right) \quad (2.4)$$

where ϑ is the set of parameters (r, θ) , P is set of points, N is total number of points and d_k is the distance of k^{th} point from the line.

- The tolerance (tol) and angle tolerance (atol) parameters specifies the allowed deviation of the parameters r and θ respectively from their optimal values.

The RAST algorithm is used with default values of these parameters. The default values of tolerance is set to 0.1 and the value of angle-tolerance is set to 0.001 in the implementation by Breuel [3].

To find lines in the astronomical images the the epsilon (eps) parameter value is set to 2 and the weight (minweight) of the line is set to 1000.

2.2.6 The Algorithm

The algorithm is used with the interest of finding the best match of all possible lines in the image. The execution steps of RAST algorithm are explained to calculate the quality function and error tolerance function. [3].

Following are the steps involved in the execution of RAST for finding lines.

- Step 1. Choose an initial region T in parameter space containing all specified parameter values.
- Step 2. Define a priority queue Q , where the priority of the queue Q is based on the upper bound on the quality of possible best match in that region as defined by the equation 2.4. The initial region is inserted into the priority queue Q .
- Step 3. Extract the element of the highest priority from the queue Q .
- Step 4. If the upper bound on the quality of the match in the extracted region T is less than the minimum quality threshold, go to step 7.
- Step 5. If the size of the region T is less than the tolerance size of the region dimensions r and θ then the region T is reported as a solution and continue to step 3.
- Step 6. If the specified size of the region T is greater than the minimum size of the interval for a solution then it is subdivided into two sub regions $T1$ and $T2$ and these sub regions are pushed into the priority queue Q and the algorithm continues to step 3.
- Step 7. Terminate

2.2.7 Output

The algorithm returns a list of parameter values r and θ . The RAST algorithm provides an output parameter *maximum result* to find more than one lines in the images.

2.3 Image Reconstruction

2.3.1 General Description

To remove lines from the image the position of the lines are computed by the RAST algorithm. To visualize the removal of the satellite tracks having geometric resemblance to lines, a new image is reconstructed from the binarized copy of the image. It is important to make a decision whether a line founded by RAST algorithm is a satellite tracks or not. This decision is based on the size of the line, that is how many pixels contribute to this line. An assumption is made that if the weight of the line is 1000 or greater, then the line found by the algorithm is considered as satellite track.

2.3.2 Painting Image

This algorithm reads the original image (PGM file) and data (line file) of pixel positions computed by the RAST algorithm. The number of lines returned by the RAST algorithm depends on the value of maxresult parameter. By setting the value of the maxresult parameter to more than one, the algorithm returns the optimal quality lines in decreasing order. The line found by the algorithm is painted black in the reconstructed image with a width of 20 pixels. The painted width of the line is decided on the basis of observed thickness of the satellite track. Examples of binarized image before and after the removal of a satellite track are shown in the following 2.5,2.6.

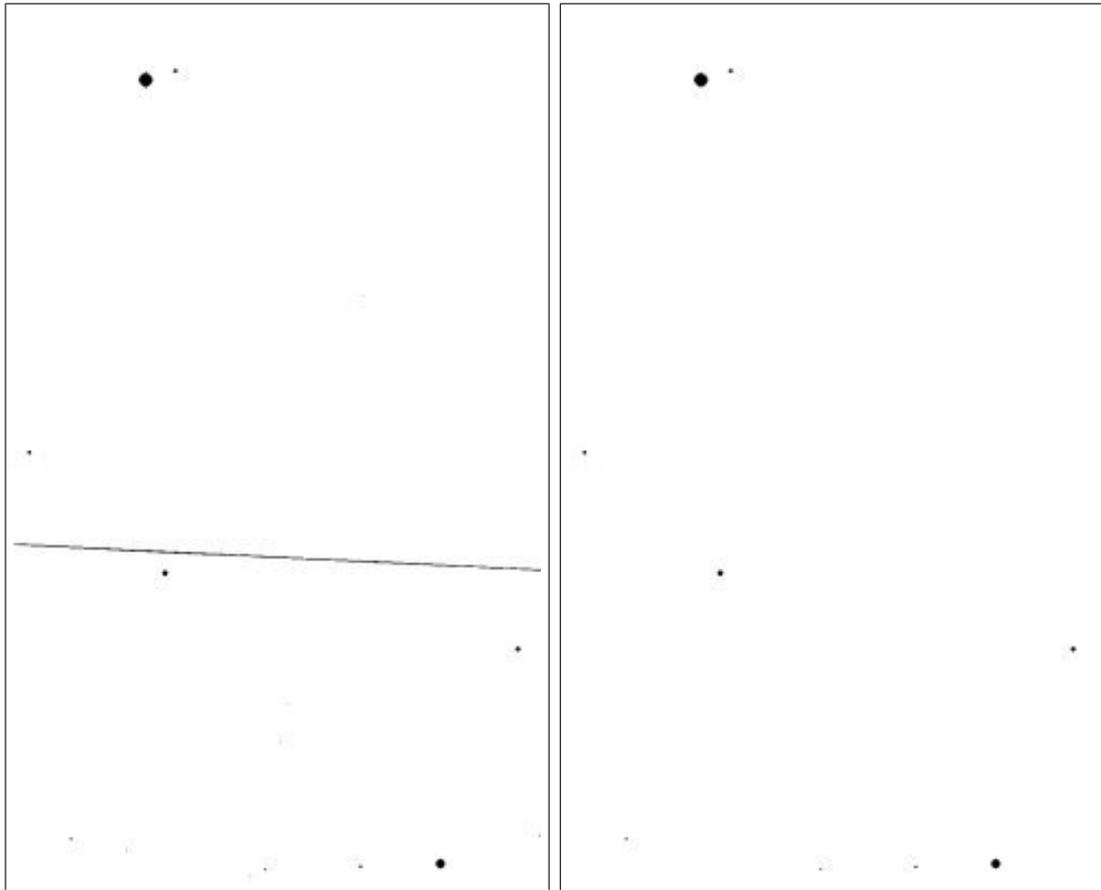


Figure 2.5: Example: Difference images of satellite tracks removal

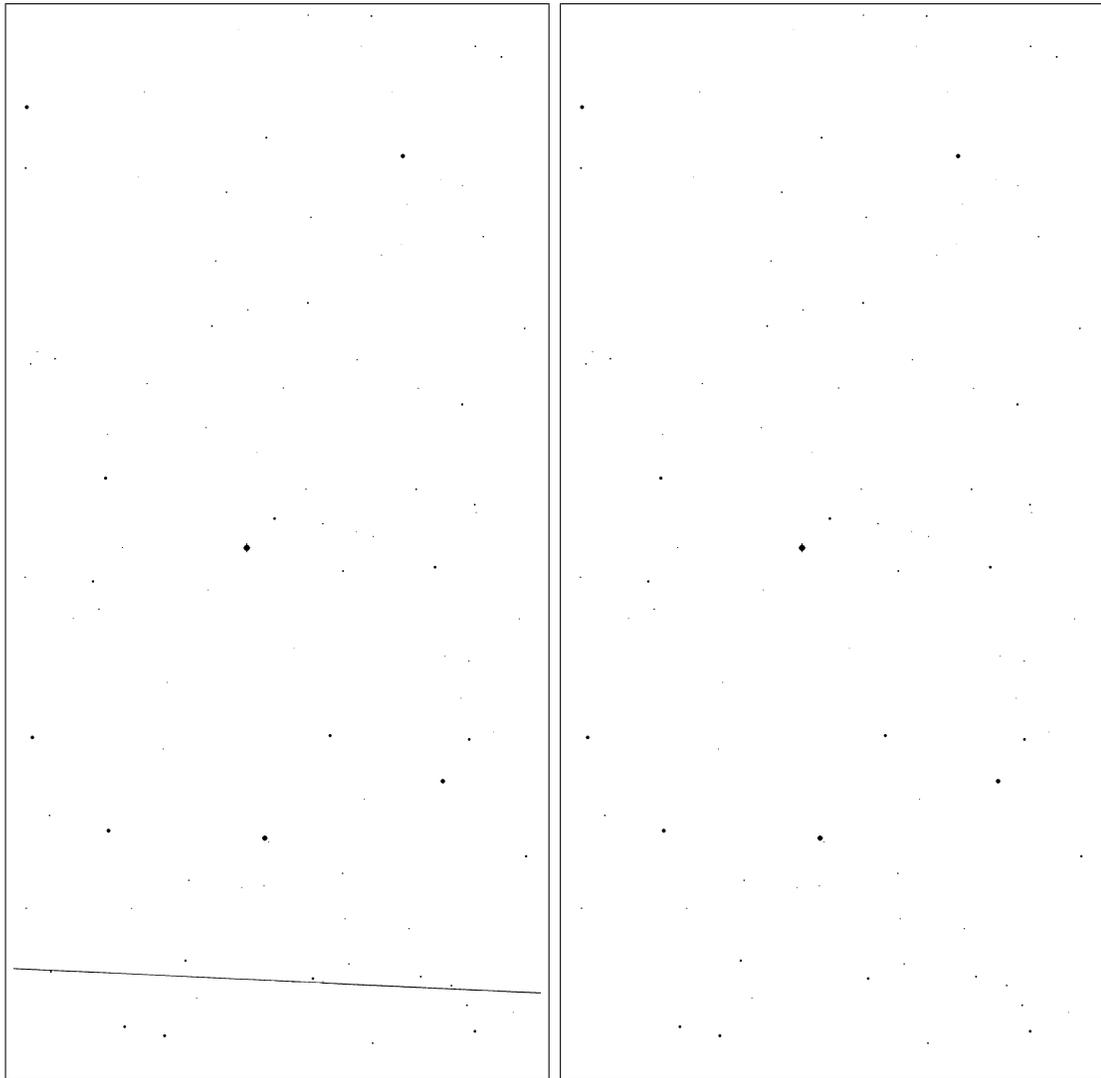


Figure 2.6: Example continued: Binarized image containing satellite track : Reconstructed image after removal of satellite track, in this image few objects are removed because they were very close to the satellite track

2.4 Summary

In the first part of the thesis a computation method for identification and removal of satellite tracks in astronomical images is presented as an alternative to the manual approach. This computation method is based on geometric matching techniques for finding lines in the images. There are different methods available for finding geometric shapes in images. As satellite tracks has geometric resemblance to lines, the focus of this part of thesis was on finding lines in the astronomical images. The RAST algorithm is applied to find these satellite tracks. The problem in this method was to find the right size of a satellite track. The problem of measurement of the size of satellite track is solved by visualizing satellite tracks of different sizes. An assumption is made that the minimum weight of the line formed by a satellite track is 1000 pixels. Results of this method are visualized by reconstructing the image after removal of satellite tracks. A numeric evaluation is given in Chapter 4 and Section 4.1. The provided solution is quite fast. It can find and remove satellite tracks from an image of size 2000x4000 pixels on a Pentium 4 processor running at 2 GHz with a typical computational time less than 60 seconds. This computation time could be decreased with high performance hardware. This solution will provide ease to astronomers for their further processing of objects in the images.

Chapter 3

Measurement of Galaxy Ellipticity

3.1 Image Pre-Processing

3.1.1 Conversion

As explained in the section 2.1, given FITS format images were converted to binary format. In addition, the FITS images were also converted to grayscale image (PGM) format. The converted grayscale and binary images were used for further processing.

3.2 Detection of Objects

After pre-processing of the binary images, the idea is to find the position of all the foreground objects in the images. The position of the foreground objects is needed for the application of "Maximum Likelihood Estimation" to measure elliptic shape galaxies.

3.2.1 Connected Components

The connected components algorithm is applied to identify the position of all the foreground objects in the image. All the white pixels in a binary image which are connected to each other define a cluster of "Connected Components". These pixels should be connected to each other in a way that a path could be established between

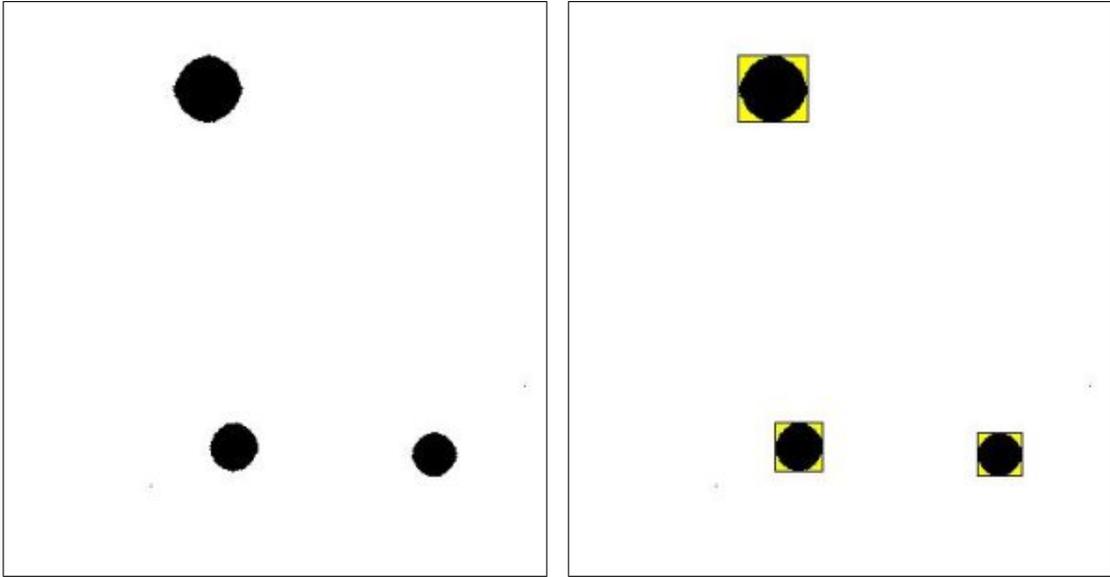


Figure 3.1: Example of bounding boxes painted as red and white spaces painted as yellow

the connected points. There are different implementations of connected components algorithm like the *Regions Growing Method* and *Labeling Method*. In this work the labeling method is used to find connected components in a binarized image as described in Chapter 4 and Section 4.1 of [7].

In order to prepare data for "Maximum Likelihood Estimation" a bounding box, defined by its coordinates x -low, x -high, y -low and y -high is computed for each cluster of connected components. The bounding box is the smallest rectangle in which the connected components fall.

To visualize the effect of connected components, the copy of binary image is constructed and rectangles are drawn in the image. These rectangles are the bounding boxes painted as red. The white spaces in a rectangle are painted as yellow. An illustration of identified connected components is presented in Figure 3.1. Only one object is assumed in a bounding box. A Gaussian grayscale distribution is estimated for each bounding box.

3.3 Measurement of Galaxies

3.3.1 Gaussian Grayscale Distribution

Gaussian grayscale distribution is calculated on the basis of sample data points. All the points in a bounding box constitute sample data. A Gaussian grayscale distribution is calculated to describe physical distribution of events.

If \mathbf{X} is the vector of sample data points then as given by [11] the probability of each sample is:

$$P(X_i) \tag{3.1}$$

Therefore the probability for whole set of sample data points based on $P(X_i)$ would be:

$$P(\mathbf{X}) = \prod_{i=1}^n \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(X_i-\mu)^T \Sigma^{-1}(X_i-\mu)} \tag{3.2}$$

The parameters mean μ and the covariance matrix Σ for a Gaussian distribution are estimated using *Maximum Likelihood Estimation*.

3.3.2 Maximum Likelihood Estimation

In the Gaussian case the maximum likelihood estimation is the process of calculating the mean and the covariance matrix. If the samples are considered as cloud of points, then the sample mean would be the centroid of the cloud which is the arithmetic average of the samples. If x_1, \dots, x_n are sample data points obtained from the distribution of x then the mean μ is defined as:

$$\mu = \frac{1}{n} \sum_{k=1}^n (x_k) \tag{3.3}$$

If x_1, \dots, x_n are the sample data points then a covariance matrix Σ is estimated as

[11]:

$$\Sigma = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)(x_k - \mu)^T \quad (3.4)$$

For astronomical images the value of a pixel represents the brightness. The elliptical galaxies have a characteristic that they are very bright at the center and faint in the boundaries. Therefore the elliptical galaxies are modeled by Gaussian grayscale distribution.

In the binary case the mean μ and the covariance matrix Σ as given by the equations (3.3, 3.4) are normalized by the total number of sample points n . In the grayscale image each sample point x_i may have a value in the range 0-255. Therefore the normalization is done by the sum of the grey values of all the sample data points x_1, \dots, x_n .

In this case $x_i = (x, y)^T$ represents the position of a foreground pixel. Let $g[x_i]$ be the intensity or value of the pixel at (x, y) then the normalization factor α for Gaussian grayscale distribution is calculated as:

$$\alpha = \sum_{x_{low}}^{x_{high}} \sum_{y_{low}}^{y_{high}} g[x, y] \quad (3.5)$$

After obtaining the scale factor α the mean μ is calculated as:

$$\mu = \frac{1}{\alpha} \sum_{x_{low}}^{x_{high}} \sum_{y_{low}}^{y_{high}} g[x, y] \begin{pmatrix} x \\ y \end{pmatrix} \quad (3.6)$$

Once the scale factor α and the mean μ are calculated then the covariance matrix Σ for Gaussian grayscale distribution is calculated as:

$$\Sigma = \frac{1}{\alpha} \sum_{x_{low}}^{x_{high}} \sum_{y_{low}}^{y_{high}} g[x, y] \left(\begin{pmatrix} x \\ y \end{pmatrix} - \mu \right) \left(\begin{pmatrix} x \\ y \end{pmatrix} - \mu \right)^T \quad (3.7)$$

The maximum likelihood estimation for unknown mean μ and covariance matrix Σ is described in Chapter 3 and Section 3.2.3 of [11]. A C++ program was written to estimate maximum likelihood by estimating parameters (mean, covariance matrix

and scale factor).

3.3.3 Image Reconstruction

To see how good this maximum likelihood estimation technique works an application is developed for visual evaluation. As the estimated parameters calculated in the previous section do not give intuitive understanding. A new synthetic image $h[x, y]$ of the size of the original image is created from the estimated parameters (mean μ , covariance matrix Σ , scale factor α and the number of samples (the dimensions of each bounding box) to visualize the result of the modeled galaxies.

Once the Gaussian grayscale distribution is calculated then it is multiplied by the scale factor α as discussed in 3.5.

$$h[x, y] = \frac{\alpha}{2\pi|\Sigma|^{\frac{1}{2}}} \exp^{-\frac{1}{2}((\begin{smallmatrix} x \\ y \end{smallmatrix})-\mu)^T \Sigma^{-1} ((\begin{smallmatrix} x \\ y \end{smallmatrix})-\mu)} \quad (3.8)$$

While reconstructing the synthetic image $h[x, y]$ each bounding box is replaced by the object constructed from the estimated parameters. These estimated parameters define the grayscale distribution which should be in the range from 0 to 255. If the value exceeds 255 then this value is set to grayscale maximum value 255. An example of Gaussian grayscale distribution is drawn in Figure 3.2.

While reconstructing the image, each pixel value of a bounding box is replaced with the estimated grayscale value. This Gaussian grayscale distribution gives the maximum intensity value of pixels at the center of the object and this intensity value decreases in the outer bound of the object. A C++ program is written to reconstruct this synthetic image from the estimated parameters. A more detailed evaluation is given in Chapter 4 and Section 4.2. Different images containing original grayscale galaxies and reconstructed galaxies are presented here.

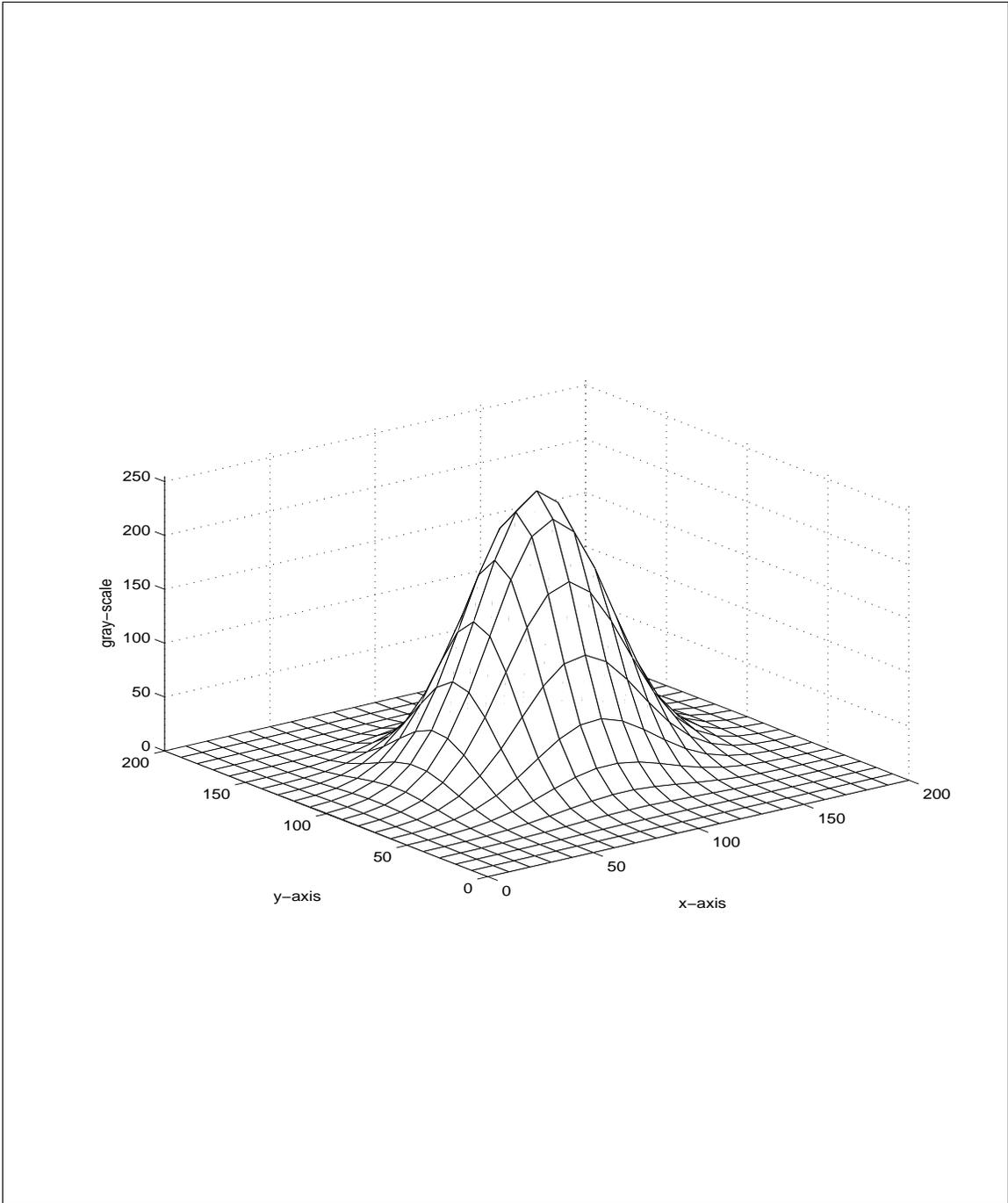


Figure 3.2: An example of Gaussian grayscale distribution

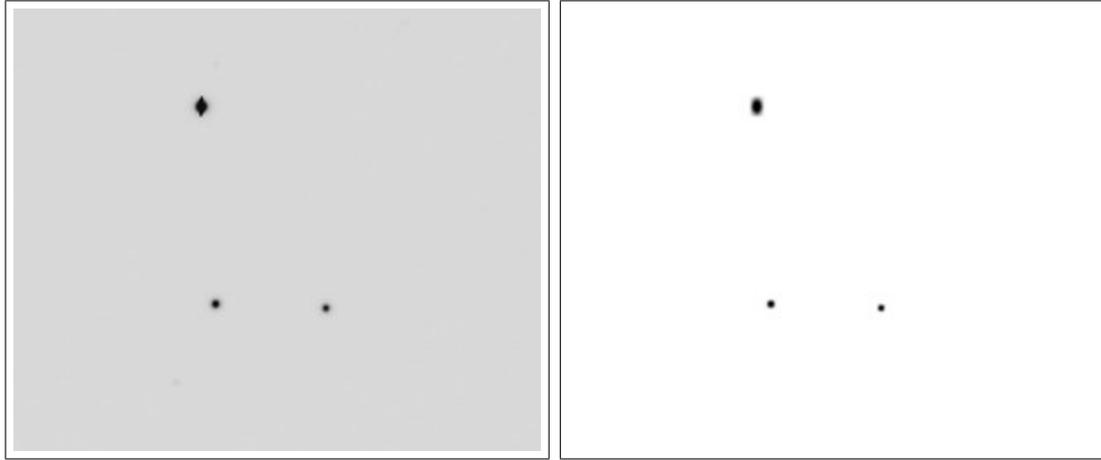


Figure 3.3: Example: Original and reconstructed galaxies

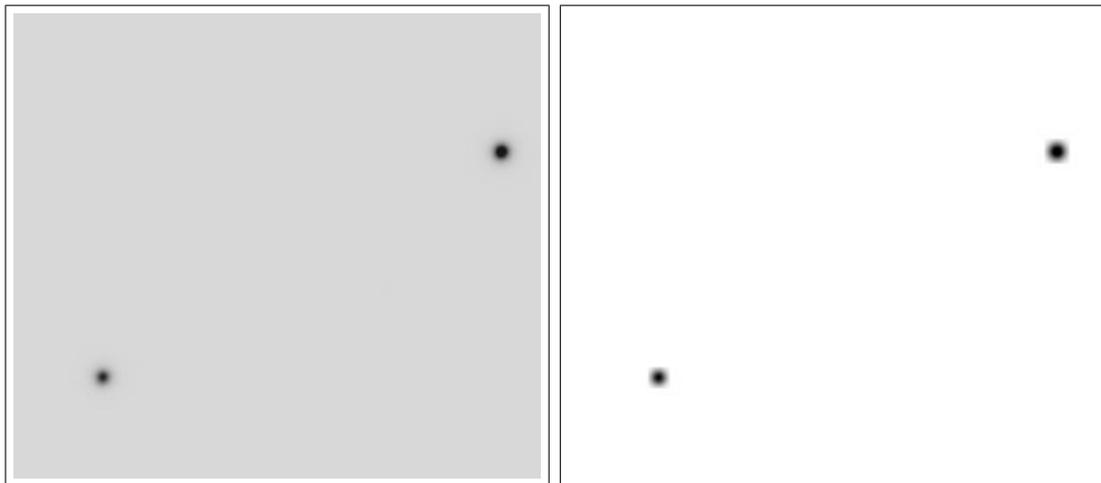


Figure 3.4: Example - Continued : Original grayscale galaxies : Reconstructed galaxies from estimate parameters

3.4 Summary

In this part of thesis work, the idea was to measure elliptic shape galaxies by estimating parameters in the astronomical images. Mostly, elliptic shape galaxies look like a scaled Gaussian. So in this work Gaussian grayscale distributions are calculated. A connected component algorithm is applied to detect the position of foreground objects in the image. This connected component algorithm uses labeling method to compute pixel positions of the objects which are connected to each other. The algorithm generates bounding boxes as output. Each bounding box is treated as one object in the image. Based on the assumption of each bounding box having one object and the shape of this object is like a scaled Gaussian the statistical method maximum likelihood estimation is applied to estimate parameters of each Gaussian. For visual examination, a synthetic image is reconstructed from the estimate parameters.

Chapter 4

Evaluation

The evaluation of two tasks - removal of satellite tracks and the measurement of galaxy ellipticity - are done separately using a large database of images provided by the Astronomy department of the University of Bonn, Germany. This database consists of 1000 images.

4.1 Removal of Satellite Tracks

A total of 102 random images are chosen to evaluate this method. To generate ground truth information, the images are visually examined and annotated whether they contain satellite tracks or not. The system described in Chapter 2 is applied to find satellite tracks in the given dataset of images.

4.1.1 Results

The results of the system are presented in table 4.1. The table gives *true positive* (T_p), *true negative* (T_n), *false positive* (F_p) and *false negative* (F_n) number of satellite tracks detection.

Table 4.1: Different types of errors made by satellite tracks detection application. Each column represents the different type of error. The column labels are: total true positive (T_p), total true negative (T_n), total false positive (F_p), total false negative (F_n)

	Total Images	T_p	T_n	F_p	F_n
Satellite Track Detection	102	2	87	5	8

True positive (T_p) are the satellite tracks which actually exists and are reported as tracks as well. True negative (T_n) are the images without satellite tracks, which are correctly separated as free of tracks. False positive (F_p) are the images which do not contain satellite tracks but are detected as images containing satellite tracks. False negative (F_n) are the images containing satellite tracks but are not reported as images containing satellite tracks.

The system has detected 2 true positive (T_p) images containing satellite tracks which are processed by the system as well. There are 87 true negative (T_n) images identified by the system where there is no satellite track exists. Therefore the accuracy of system for finding satellite tracks is 87.3%, which is the total number of correct detections.

The system has detected 5 false positive (F_p) images. The false positive (F_p) images are analyzed and it turned out to some expanded stars and galaxies in the image due to the brightness and make a *Light Spike* in the available database of images. These light spikes look like vertical satellite tracks. While considering these light spikes as satellite tracks the algorithm detects and removes them as well. These light spikes are removed completely except the circular shaped bright stars or galaxies due to their thickness. An illustration of the images having stars and galaxies effected with the telescope light spikes before and after removal of these light spikes are presented in the Figure 4.1.

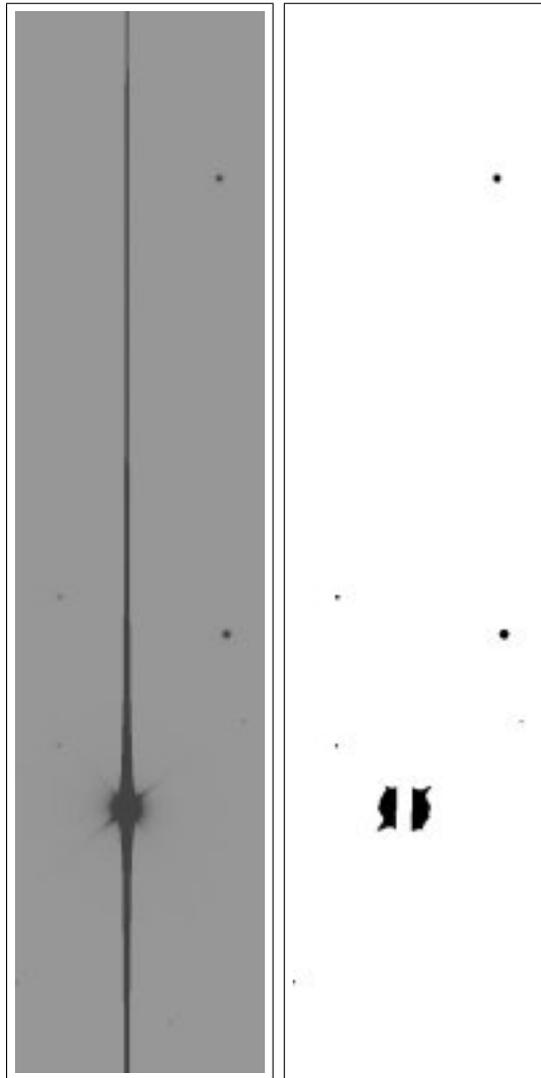


Figure 4.1: The original image containing light spikes, there after the removal of these light spikes a few points of the galaxies still remains visible in the output image

Table 4.2: Different types of results made by RAST algorithm while finding satellite tracks in the images. Each column represents different type of error. The column labels are: true positive (T_p), true negative (T_n), false positive (F_p), false negative (F_n)

Algorithm	Total Images	T_p	T_n	F_p	F_n
RAST	94	2	87	5	0

The system has detected 8 False negative (F_n) images. These False negative (F_n) images are analyzed and the images were found to contain satellite tracks but in very low brightness. The system was not able to detect satellite tracks in them. After analyzing False negative (F_n) images deeply, it is found that the original FITS images do not contain *DATAMIN* and *DATAMAX* fields. As the header data units (HDUs) of FITS images have *DATAMIN* and *DATAMAX* containing the range of grayscale values in the images. *DATAMIN* describes the minimum and *DATAMAX* describes the maximum gray value in the image. The foreground objects are lost already in the conversion step, so there is no chance of detection by RAST algorithm. This failure was not because of the RAST algorithm but was due to the missing information about *DATAMIN* and *DATAMAX* in FITS header data units (HDUs) which are the essential information required for the conversion process.

Therefore a new evaluation scheme was decided where these black images were removed from the test set. The results of this new evaluation scheme is presented in a separate table 4.2.

As the table shows that the RAST algorithm did not find any false negative (F_n) results. The accuracy of our algorithm is increased to 94.7% with this evaluation scheme.

4.2 Measurement of Galaxy ellipticity

As there is no ground truth available which provides the parameters of the elliptical galaxies. So, without the ground truth a numerical evaluation of the maximum-likelihood estimation could not be conducted. To see how good the maximum likelihood estimation technique is in estimating parameters for elliptic shape galaxies, a subjective evaluation scheme is conducted. The evaluation of this method is organized by comparing the original image with the reconstructed synthetic image.

4.2.1 Results

Both the images (original and modeled) are examined and the results are analyzed visually. Based on the subjective evaluation scheme results of this method is organized in two categories: elliptic shape modeled galaxies and error modeled shapes.

While evaluating the modeled shapes and comparing them with the original images different elliptic shape modeled galaxies were found. Some of the elliptic shape modeled galaxies with different rotations are presented in the Figures 4.2(a) to 4.3(c).

While evaluating the modeled shapes and comparing them with the original images different error modeled shapes were found. Some of the error modeled shapes are presented in the Figures 4.4 to 4.5.

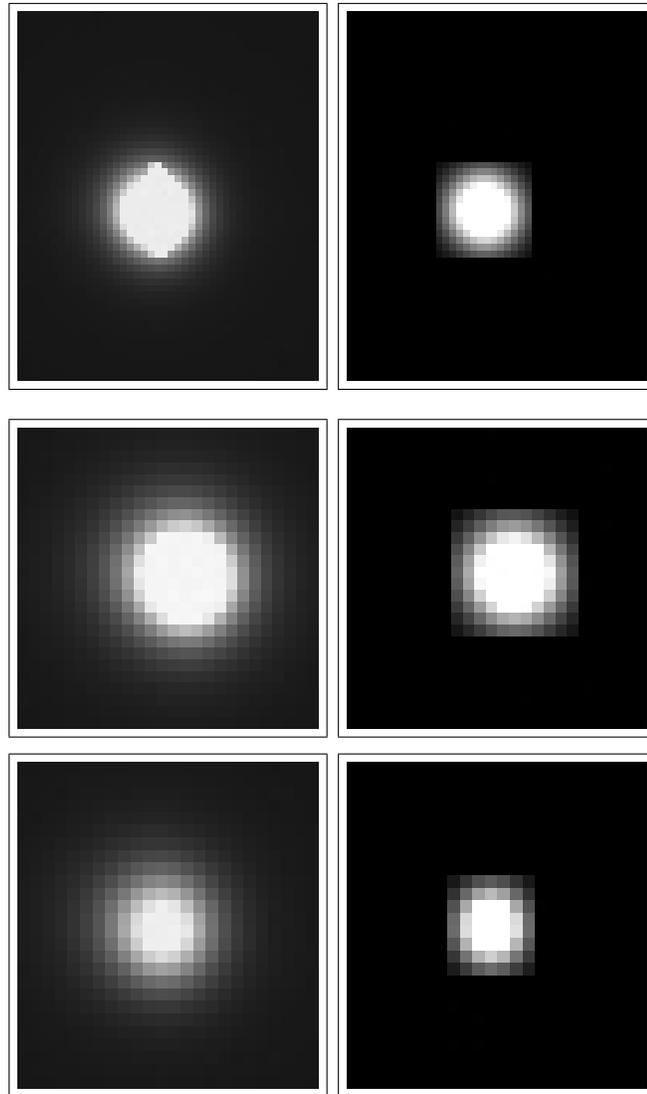


Figure 4.2: Example of galaxies reconstruction

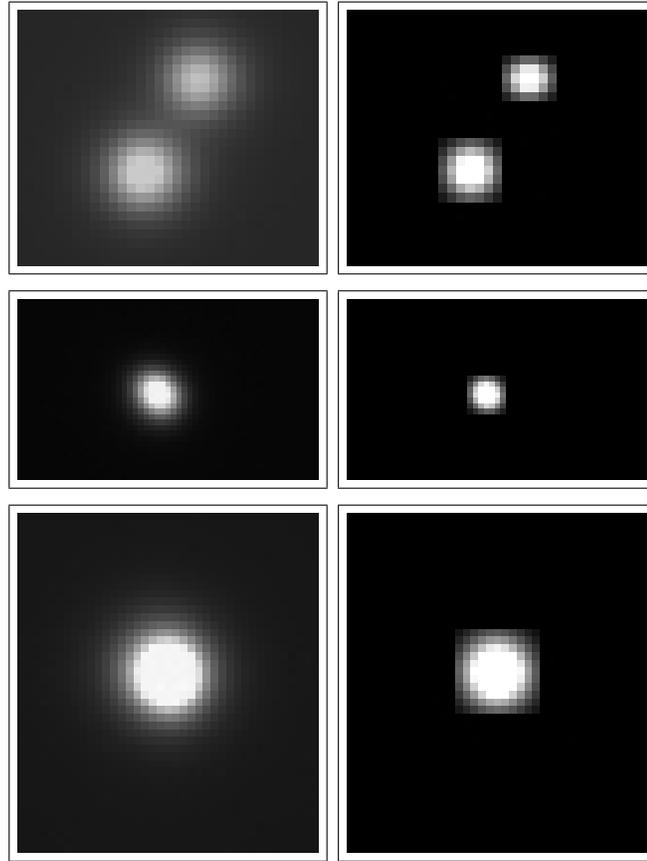


Figure 4.3: Example - continued: On the left hand side in the original grayscale images there are different elliptical galaxies and on the right hand side there are elliptic shape galaxies. These reconstructed shapes of the galaxies with different rotation are similar to the elliptic shape galaxies in the original images

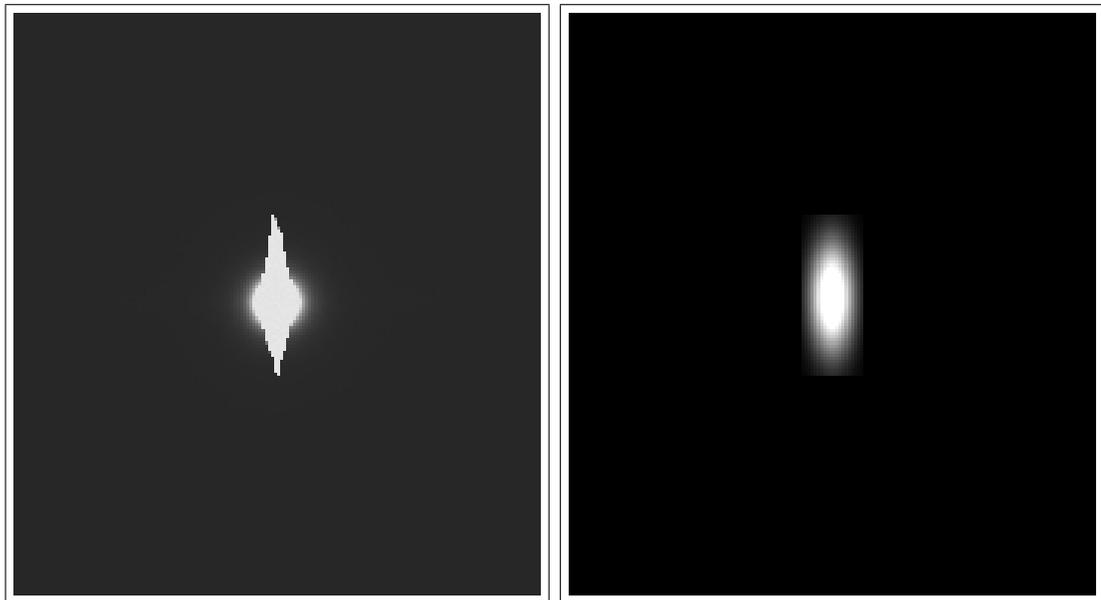


Figure 4.4: Example: On the original grayscale image there is a galaxy with telescope light spike which look like an expanded galaxy. In the reconstructed image the error shape is modeled as an elliptic shape galaxy

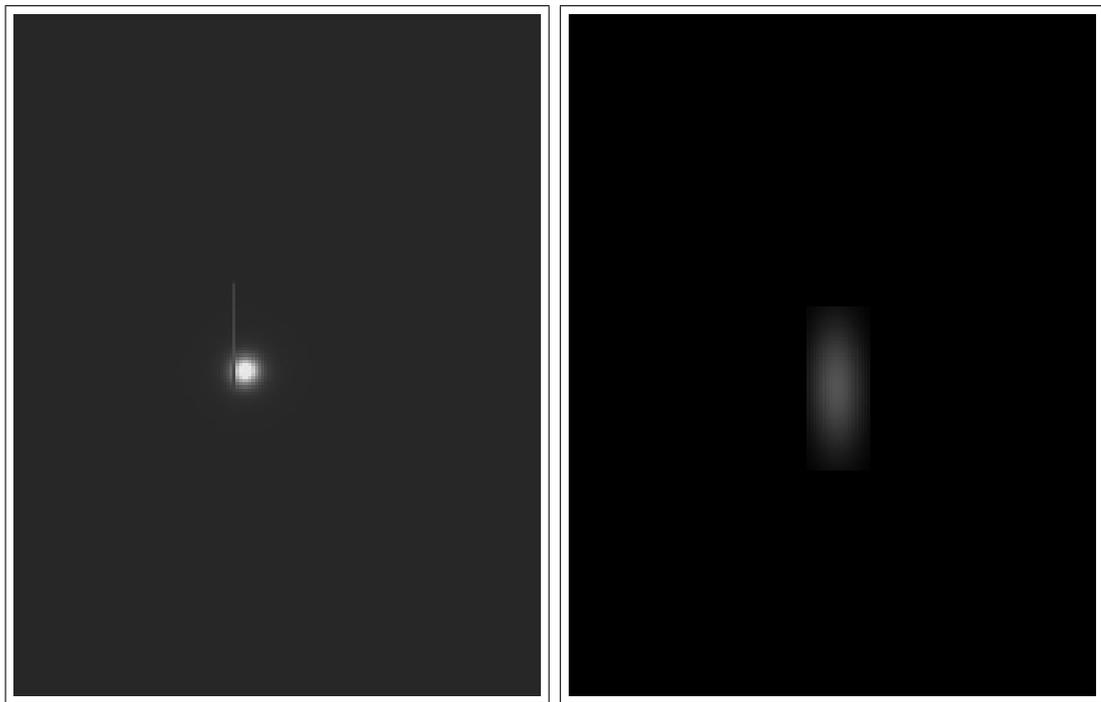


Figure 4.5: Example - continued: On the original grayscale image there is a galaxy with CCD-defect form a small line. In the reconstructed image the error shape is modeled as an elliptic shape galaxy

Chapter 5

Conclusion and Future Work

We have developed a system to find satellite tracks and to measure elliptic shape galaxies in the images taken by astronomers. In this system the removal of satellite tracks is done using the geometric matching algorithm RAST. While identifying the satellite tracks in astronomical images, it is important to know the actual size of the satellite tracks. Currently, the system is implemented based on the observation that if 1000 or more pixels contribute to make a line then this line is a satellite track. The numerical evaluation of this part of the system is done on a large database of astronomical images and the performance of this system is presented. The accuracy of our system is 94.7%. The second part of this system estimates different parameters in order to measure galaxy ellipticity using a statistical method Maximum Likelihood Estimation. We have presented a subjective evaluation of this part of the system.

This system can be extended in several ways.

- Currently, the observed thickness of the satellite tracks is set to width of 20 pixels which could be refined after knowing the minimum size of the satellite track.
- Currently, every object in a bounding box is modeled as an elliptic shape and considered as a single scaled Gaussian. But there could be two or more overlapping objects in each bounding box. To solve this problem, a *Gaussian Mixture*

Model could be applied to classify overlapping objects.

- Currently, parameters of the elliptic shape galaxies are estimated and their shapes are modeled. After estimating other shapes of galaxies like *Normal Spirals* and *Barred Spirals* etc. they could be further classified.

Bibliography

- [1] E. Bertin. Sextractor, users manual. <http://terapix.iap.fr/soft/sextractor>, 2005.
- [2] T.M. Breuel. *Geometric Aspects of Visual Object Recognition*. PhD thesis, Massachusetts Institute of Technology, 1992.
- [3] T.M. Breuel. Finding lines under bounded error. *Pattern Recognition Letters*, 29(1):167–78, 1996.
- [4] T.M. Breuel. A practical, globally optimal algorithm for geometric matching under uncertainty. In *International Workshop on Combinatorial Image Analysis (IWCIA August 2001), Philadelphia, CA*, 2001.
- [5] Dr T. Erben. Personal communication, 2005.
- [6] T. Erben and M. Schirmer et. al. Gabods: The garching-bonn deep survey; iv. methods for the image reduction of multi-chip cameras. arXiv.org:astro-ph/0501144, 2005.
- [7] Berthold K.P. Horn. *Robot Vision*. MIT Press, 2001.
- [8] R. Nemiroff and J. Bonnell. Galaxies. <http://antwrp.gsfc.nasa.gov>, 2006.
- [9] Dr. William D. Pence. Fits overview. <http://fits.gsfc.nasa.gov>, 2004.
- [10] R. E. Woods R. C. Gonzalez. *Digital Image Processing, 2nd Edition*. Prentice-Hall, Inc., 2002.
- [11] D. G. Stork R. O. Duda, P. E. Hart. *Pattern Classification, 2nd Edition*. John Wiley and Sons, Inc., 2001.

- [12] Alexandre Refregier, Tzu-Ching Chang, and David Bacon. A new method to measure galaxy shapes. [arXiv.org:astro-ph/0202023](https://arxiv.org/abs/2020.02.23), 2002.
- [13] A.J Storkey, Hambly N.C., C.K.I. Williams, and R.G. Mann. Cleaning sky survey databases using Hough Transform and renewal string approaches. *Monthly Notices of the Royal Astronomical Society*, 347(1):36–51, 2004.