Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

DEVELOPMENT OF ADVANCED ALGORITHMS TO DETECT, CHARACTERIZE AND FORECAST SOLAR ACTIVITIES

by

Yuan Yuan

Study of the solar activity is an important part of space weather research. It is facing serious challenges because of large data volume, which requires application of state-of-the-art machine learning and computer vision techniques. This dissertation targets at two essential aspects in space weather research: automatic feature detection and forecasting of eruptive events.

Feature detection includes solar filament detection and solar fibril tracing. A solar filament consists of a mass of gas suspended over the chromosphere by magnetic fields and seen as a dark, ribbon-shaped feature on the bright solar disk in Hα (Hydrogen-alpha) full-disk solar images. In this dissertation, an automatic solar filament detection and characterization method is presented. The investigation illustrates that the statistical distribution of the Laplacian filter responses of a solar disk contains a special signature which can be used to identify the best threshold value for solar filament segmentation. Experimental results show that this property holds across different solar images obtained by different solar observatories. Evaluation of the proposed method shows that the accuracy rate for filament detection is more than 95% as measured by filament number and more than 99% as measured by filament area, which indicates that only a small fraction of tiny filaments are missing from the detection results. Comparisons indicate that the proposed method outperforms a previous method. Based on the proposed filament segmentation and characterization method, a filament tracking method is put forward,

which is capable of tracking filaments throughout their disk passage. With filament tracking, the variation of filaments can be easily recorded.

Solar fibrils are tiny dark threads of masses in Hα images. It is generally believed that fibrils are magnetic field-aligned, primarily due to the reason that the high electrical conductivity of the solar atmosphere freezes the ionized mass in magnetic field lines and prevents them from diffusing across the lines. In this dissertation, a method that automatically segments and models fibrils from Hα images is proposed. Experimental results show that the proposed method is very successful to derive traces of most fibrils. This is critical for determining the non-potentiality of active regions.

Solar flares are generated by the sudden and intense release of energy stored in solar magnetic fields, which can have a significant impact on the near earth space environment (so called space weather). In this dissertation, an automated solar flare forecasting method is presented. The proposed method utilizes logistic regression and SVM (support vector machine) to forecast the occurrences of solar flares based on photospheric magnetic features. Logistic regression is used to derive the probabilities of solar flares occurrence, which are then fed to SVM for determining whether a flare will occur. Comparisons with existing methods show that there is an improvement in the accuracy of X-class solar flare forecasting. It is also found that when sunspot-group classification is combined with photospheric magnetic parameters, the performance of flare forecasting can be further lifted.

DEVELOPMENT OF ADVANCED ALGORITHMS TO DETECT, CHARACTERIZE AND FORECAST SOLAR ACTIVITIES

by
Yuan Yuan

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2011

.

APPROVAL PAGE

DEVELOPMENT OF ADVANCED ALGORITHMS TO DETECT, CHARACTERIZE AND FORECAST SOLAR ACTIVITIES

Yuan Yuan

Dr. Frank Y. Shih, Dissertation Co-Advisor Professor of Computer Science, NJIT	Date
Dr. Ju Jing, Dissertation Co-Advisor Research Professor of Physics, NJIT	Date
Dr. Haimin Wang, Dissertation Co-Advisor Distinguished Professor of Physics, NJIT	Date
Dr. Alexandros V. Gerbessiotis, Committee Member Associate Professor of Computer Science, NJIT	Date
Dr. Cristian M. Borcea, Committee Member Associate Professor of Computer Science, NJIT	Date

BIOGRAPHICAL SKETCH

Author: Yuan Yuan

Degree: Doctor of Philosophy

Date: May 2011

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2011
- Master of Engineering in Computer Applied Technology,
  Zhejiang University of Technology, Hangzhou, P. R. China, 2008
- Bachelor of Engineering in Computer Science and Technology,
  Zhejiang University of Technology, Hangzhou, P. R. China, 2004

Major: Computer Science

Presentations and Publications:

Yuan, Y., Shih, F. Y., Jing, J., Wang, H. and Chae, J., "Automatic solar filament segmentation and characterization," accepted by Solar Physics, 2011.

Jing, J., Yuan, Y., Wiegelmann, T., Xu, Y., Liu, R. and Wang, H., "Nonlinear force-free modeling of magnetic fields in a solar filament." The Astrophysical Journal Letters, vol. 719, pp. L56-L59, August 2010.

Yuan, Y., Shih, F. Y., Jing, J. and Wang, H., "Automated flare forecasting using a statistical learning technique," Research in Astronomy and Astrophysics, vol. 10, no. 8, pp. 785-796, July 2010.

Jing, J., Tan, C., Yuan, Y., Wang, B., Wiegelmann, T., Xu, Y. and Wang, H., "Free magnetic energy and flare productivity of active regions," The Astrophysical Journal, vol. 713, pp. 440-449, April 2010.

Shih, F. Y. and Yuan, Y., "A wavelet-based encoding algorithm for high dynamic range images," The Open Signal Processing Journal, vol. 3, pp. 13-19, 2010.

Shih, F. Y. and Yuan, Y., "A comparison study on copy-cover image forgery detection," The Open Artificial Intelligence Journal, vol. 4, pp. 49-54, 2010.

Yuan, Y., Shih, F. Y., Jing, J. and Wang, H., "Solar flare forecasting using sunspot-group classification and photospheric magnetic parameters," IAU Symposium 273: Physics of Sun and Star Spots, Ventura, California 22-26 August 2010.

Yuan, Y., Shih, F. Y., Jing, J. and Wang, H., "Solar filament extraction and characterizing," 216th American Astronomical Society Meeting, Miami, FL, May 2010.

Jing, J., Tan, C., Yuan, Y., Wang, B., Wiegelmann, T., Xu, Y. and Wang, H., "Free magnetic energy and flare productivity of active regions," 216th American Astronomical Society Meeting, Miami, FL, May 2010.

Yuan, Y., Shih, F. Y. and Jing, J., "An automated flare forecasting system using statistical and machine learning techniques," American Astronomical Society, Solar Physics Division Meeting, Boulder, CO, June 2009.

To my family,
For their endless love and support.

ACKNOWLEDGMENT

My most sincere thanks go to my dissertation advisors Dr. Frank Y. Shih, Dr. Ju Jing and Dr. Haimin Wang for introducing me to the wonders of scientific research. I thank them for their guidance, encouragement and support during the past five years.

I am grateful to my committee members: Dr. Alexandros Gerbessiotis and Dr. Cristian M Borcea. They have made lots of thoughtful suggestions for improvement.

I am deeply thankful for the financial support from Dr. Jongchul Chae of Seoul National University. I also would like to thank Dr. Thomas Wiegelmann of Max-Planck-Institut für Sonnensystemforschung for his helpful comments.

I want to express my sincere gratitude to past and present members of Space Weather Research Lab and Computer Vision Lab at NJIT. I thank Dr. Dale E. Gary, Dr. Wenda Cao, Dr. Jeongwoo Lee, Dr. Gang Fu, Dr. Ming Qu, Dr. Yan Xu, Dr. Changyi Tan, Dr. Rui Liu, Dr. Zhiwei Liu, Dr. Chang Liu, Dr. Na Deng, Dr. Sung-Hong Park, Yixuan Li, Shuo Wang, Xin Chen, Xiupeng Wang, Zhicheng Zeng, Faizan H Naqvi, Chandralekha De, Venkata Gopal Edupuganti and Jinwen Liu.

Special thanks should be given to Christine A. Oertel and Angel J. Bell for assisting me in many ways so as to make my study easier.

Last but not the least, I would like to thank my family for their endless love, encouragement and support. My gratitude to them can never be overstated.

The research work presented in this dissertation is supported by the National Science Foundation (NSF) under grants ATM 09-36665, ATM 07-16950, ATM-0745744 and National Aeronautics and Space Administration (NASA) under grants NNX0-7AH78G, NNXO-8AQ90G.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

LIST OF FIGURES
(Continued)

LIST OF ABBREVIATIONS

BBSO	Big Bear Solar Observatory
CME	Coronal Mass Ejections
GPS	Global Positioning System
Hα	Hydrogen Alpha
KANZ	Kanzelhöhe Solar Observatory
MDI	Michelson Doppler Imager
Mm	Megameter
Mm^2	Square Megameter
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
NSF	National Science Foundation
OACT	Catania Astrophysical Observatory
SOHO	Solar and Heliospheric Observatory
SVM	Support Vector Machine
YNAO	Yunnan Astronomical Observatory

CHAPTER 1

INTRODUCTION

1.1 Overview

Space weather refers to the conditions in the space environment caused by the Sun that can affect the performance and reliability of space-borne and ground-based system as well as human life and health [1]. Space weather can significantly impact satellite communications and navigation, interrupt the Global Positioning System (GPS), shorten orbital lifetime of Earth-orbiting satellites, damage satellites' electronic components, destroy electric power system distribution grids, bring radiation sickness to astronauts, and endanger passengers on commercial flights going through polar route [2].

To accurately predict space weather, it is required to understand the Sun and especially the mechanism of eruptive events on the Sun, like solar flares and Coronal Mass Ejections (CMEs). Solar flares are due to the sudden and intense release of energy stored in solar magnetic fields [3]. CMEs are the most energetic events in the solar system, during which coronal material of mass up to 10^{16} g is expelled at speeds of several 10^{2} to 10^{3} kilometers per second from the Sun [1]. CMEs are associated with solar flares and filament eruption (whole or part of filament ascends with velocity of several hundred km/s [4]). To understand the mechanism of solar flares and solar filaments is of vital importance to solar physics.

Rapid progress of technology makes it possible to establish both space-borne and ground-based solar observatories, which provide higher quality and larger quantity of solar images than ever before. It brings both hope and challenges. On the one hand, more and more data with better quality are readily available; on the other hand, the tools to crunch the

data to derive meaningful results from the data are not available. For example, the Global High-Resolution H-alpha Network [5] currently collect H-alpha full disk solar images from eight different solar observatories across the world, each of which produces dozens of images per day. The recently launched satellite SDO (Solar Dynamics Observatory) produces 4TB data per day. It is not practical to measure the features (like solar filaments) on solar disk manually.

Primary goal of this study is to develop automated tools that can detect solar features (filaments) and predict solar eruptive events (flares) using advanced digital image processing and machine learning techniques.

1.2 Solar Filament Segmentation and Fibril Tracing

Solar filaments (also called prominence when it appears at the solar limb) are clouds of relatively cool and dense gas suspended above the solar photosphere, generally along a magnetic neutral line [3]. Researchers are exploring the close relationship between erupting filaments and coronal mass ejections (coronal mass ejection is a high speed outward ejection of a large volume of magnetized solar plasma [4]) by studying the evolution of solar filaments [3, 6-8].

There have been several studies targeted to solar filaments detection and characterization. An automatic solar filament detection system was firstly developed by Gao *et al*. [9], in which filament detection is accomplished by segmentation using 50% of median value of the solar disk followed by region growing. This method is surpassed by Shih and Kowalski [10], which utilized localized segmentation and mathematical morphology. Bernasconi *et al*. [11] developed a solar filament segmentation and characterization technique which delves into the details of each piece of filament, such as

detection the barbs of filament. Qu *et al*. [12] adopted image enhancement, localized segmentation, morphology processing and edge linking for solar filament segmentation. Based on the comparison in [12], the method outperforms those proposed before [9, 10].

However, most proposed methods on solar filaments detection are targeted for the solar images obtained by a specific solar observatory, and thus it is relatively easier to tackle. In this study, a generic solar filament detection and characterization method is developed. Here, generic means that the method can be used for solar images obtained by different solar observatories with different statistical properties. The method deals with three aspects of solar filament segmentation. Firstly, identify the center location and radius of solar disks and then segment solar disks from solar images. Secondly, after removing the unbalanced luminance of solar disks, segment solar filaments from solar disks. Thirdly, characterize the length, location and orientation of each piece of solar filament.

Experimental results illustrate that the accuracy measured by filament area is 99% and accuracy measured by filament number is 93%. Comparison with existing methods also demonstrates the superiority of the proposed method.

Solar fibrils seen in the Hα central line are threads of mass that appear in abundance throughout the field-of-view (FOV) of Hα filtergrams. It is generally believed that fibrils are magnetic field-aligned, primarily due to the reason that the high electrical conductivity of the solar atmosphere “freezes" the ionized mass in magnetic field lines and prevents them from diffusing across the lines. Very recently, [13] tested this common notion for the first time by comparing the orientation of fibrils to the azimuth of chromospheric magnetic fields obtained by spectropolarimetric measurements of Ca II lines, and found a general alignment as well as some discrepancy between the two directions. [13] ascribed the

discrepancy to either the difference in formation height or the time lag between the fibril and magnetic field measurements.

A fibril segmentation and modeling method is presented in this dissertation. Since it is mostly true that fibrils are oriented along the magnetic field direction theoretically and observationally, it would be reasonable to adopt fibrils as a surface tracer of chromospheric magnetic fields, which helps in our understanding of the energy storage and release mechanism of solar eruptive events.

Image processing techniques such as image enhancement, image segmentation, and union-find are used to segment fibrils from Hα images. Least squares curve fitting is used to model segmented fibrils. Experimental results show that the proposed method is very successful in segmentation and modeling of most fibrils, especially major fibrils.

1.3 Solar Flare Forecasting

Solar flares are large explosions in the solar atmosphere, which typically release the order of 10^{25} joules of energy [14]. Most Solar flares can be observed as a local brightening in the Hα line [1]. According to the peak intensity of soft x-ray emission in the 0.1-0.8 nanometer band measured by Earth-orbiting satellites, solar flares can be classified into A, B, C, M and X classes [1]. Since flares below C class are in general too weak to bring major space weather events, most attention is paid to C, M and X-class flares.

It is believed that solar flares are due to the magnetic reconnection [3, 14-17]. There are many studies on the correlation between solar flares and solar magnetic properties [18-20].

Based on the statistical correlation between solar flares and solar magnetic field measures, several solar flare forecasting methods have been developed. Georgoulis and

Rust [21] developed a method of quantitative forecasting of M-class and X-class flares based on a single metric defined as the effective connected magnetic field. Barnes and Leka [22] adopted discriminant analysis to perform probabilistic forecasting of solar flares from vector magnetic field parameters. Combining the support vector machine (SVM) and the K-Nearest Neighbors (KNN), Li *et al.* [23] developed a flare forecasting model to predict whether an M-class flares will occur for each active region within two days. Song *et al*. [24] used the logistic regression as a forecasting model to estimate the probability for each active region to produce X-, M- or C-class flares. The comparison made by Song *et al*. [24] demonstrated that the proposed method outperforms those by Solar Data Analysis Center (SDAC) and NOAA's Space Weather Prediction Center (SWPC). However, there is a problem with this method, in which the predicted probability of X-class flare is underestimated.

As part of the dissertation, an automatic solar flare forecasting technique is presented, which can predict the occurrence of C, M, and X-class solar flares based on photospheric magnetic parameters. The method utilizes logistic regression and support vector machine (SVM). From an active region, magnetic parameters are extracted, and fed to a trained logistic regression model. The output of the logistic regression model (four probabilities) is further fed to a trained SVM to get the final forecasting results.

Experimental results, from a sample of 230 active regions between 1996 and 2005, show the accuracies of a 24-hour flare forecast to be 0.86, 0.72, 0.65 and 0.84 respectively for the four different classes. Comparison with the method proposed in [24] shows an improvement in the accuracy of X-class flare forecasting.

1.4 Digital Image Processing

Most of the information about the Sun used in this study is derived from digital solar images obtained by ground-based or airborne solar observatories. In this section, an introduction on basic digital image processing is presented to facilitate the understanding of the following chapters.

An image can be represented as a two-dimensional function $f(x, y)$, where x and y are spatial coordinates, and the value of f at a pair of coordinates (x, y) is called the intensity of the image at that location. When x, y and the value of f are all finite, discrete quantities, $f(x, y)$ is referred to as a digital image. A digital image is composed of a finite number of elements, each of which is located at a particular location and has a value. These elements are referred to as pixels. Digital image processing refers to processing a digital image by digital computing devices.

Digital image processing is concerned with the study and the implementation of methods for formation, communication, enhancement, and analysis of digital images. Digital image processing has been applied to a variety of fields, including astronomy [25] (telescopes), geophysics [26] (electromagnetic imaging), medical science [27] (CT, MRI, ultrasound imaging, microscopes imaging), mass communication and publishing industry [28] (printing, scanning, photocopying), entertainment [29] (special effect in movies, video games), security and digital right management [30] (digital watermarking, biometrics) and so on.

Digital image processing is composed of three basic operations, namely point operation, local operation and global operation. In point operation, the output intensity of a pixel is dependent only on the input intensity of the same pixel. In local operation, the

output intensity of a pixel is dependent on the input intensities in the neighborhood of the pixel. In global operation, the output value of a pixel depends on all the pixel intensities in the input image.

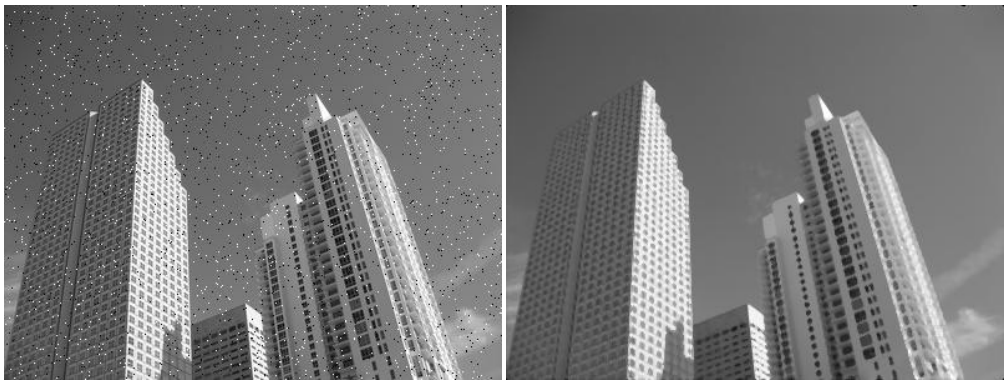
Figure 1.1 Illustration of image rotation. The image on the right panel is the result of rotation of 30 degree clockwise.

Basic operations of digital image processing can also be classified into algebraic operation, geometric operations and noise filtering. Algebraic operations include addition, subtraction, multiplication and division of digital images. Geometric operations mean to change spatial relationships between objects within an image. For example image rotation shown in Figure 1.1 is one kind of geometric operations of digital image. Noise may be introduced during image acquisition (electric noise introduced by digitizer) or transmitting (satellite images). Well-known noise filtering techniques includes Gaussian filtering and median filtering.

1.4.1 Image Filtering

Median filtering belongs to a local operator, which looks at the nearby neighbors of a pixel in the input image to determine the output intensity of the pixel. The output intensity of each pixel is the median value of its nearby neighbors of the pixel in the input image. The

median can be figured out by first sorting all the pixel intensities from the surrounding neighborhood and then picking the middle pixel intensity. Figure 1.2 illustrates an example of removing artificially introduced salt and pepper noise by median filtering.



(a) Original image (b) Image after median filtering

Figure 1.2 Illustration of median filtering to remove salt and pepper noise.

Gaussian filtering is another local operation which can be implemented with convolution. Given a kernel matrix $g(x,y)$, convolution between a digital image $f(x,y)$ and $g(x,y)$ is defined as the following equation:

$$f(x,y)*g(x,y)=\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}f(i,j)g(x-i,y-j) \tag{1.1}$$

Gaussian filtering is the convolution between a digital image $f(x,y)$ and a Gaussian kernel matrix. A two-dimensional Gaussian kernel matrix is defined as following [29] :

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1.2}$$

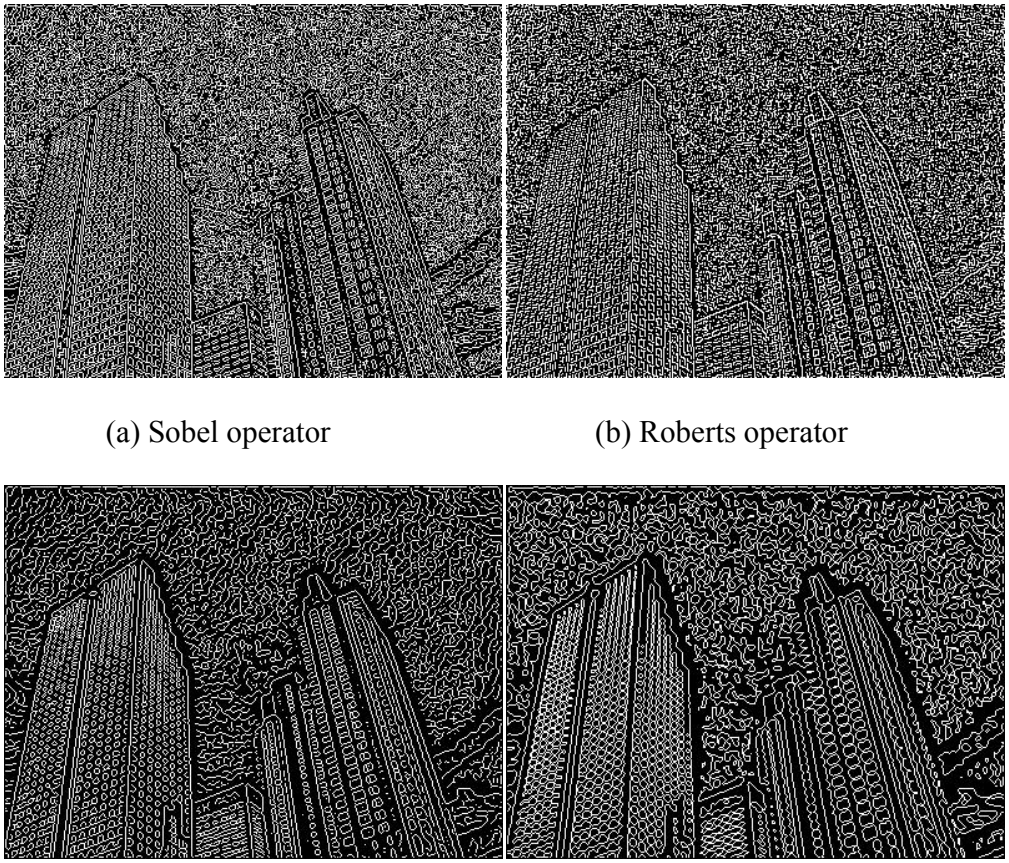
Table 1.1 Discrete Approximation to a 5×5 Gaussian Kernel with $\delta=2.5$.

0.0285	0.0363	0.0393	0.0363	0.0285
0.0363	0.0461	0.0500	0.0461	0.0363
0.0393	0.0500	0.0541	0.0500	0.0393
0.0363	0.0461	0.0500	0.0461	0.0363
0.0285	0.0363	0.0393	0.0363	0.0285

(a) Original image (b) Image after Gaussian filtering

Figure 1.3 Illustration of Gaussian filtering.

Table 1.1 illustrates a discrete approximation to a 5×5 Gaussian kernel with $\delta=2.5$. Figure 1.3 illustrates an example of Gaussian filtering, in which the image in the left panel is convoluted with the Gaussian kernel shown in Table 1.1 to produce a blurred image shown in the right panel.



(a) Sobel operator (b) Roberts operator

(c) Canny operator (d) LoG operator

Figure 1.4 Illustration of edge detection by convolution with edge operators.

.

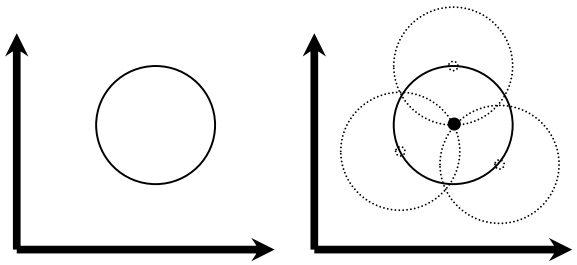
Convolution with different kernel matrix would produce different effects which is valuable in digital image processing. Among the various applications of convolution is edge detection. Edge points are located at the coordinates where the gradient magnitudes are comparatively large. Edge operators are kernel matrices, when convolution with a digital image, are able to enhance the intensities of edge points and suppress the intensities of non-edge points. Mostly well-known edge operators are Roberts operator, Sobel operator, LoG operator, and Canny operator [29]. Figure 1.4 illustrates the result of edge detection by convolution with edge operators.

1.4.2 Hough Circle Transform

Digital image processing is not only able to enhance an image (such as denoising [31]) but also able to extract higher-level information from the image. The Hough transform [32] is a feature extraction technique used to find a parameterized shape or structure from digital image processing. Hough circle transform is used in this study to identify the radius and center location of the solar limb in a solar image. A solar limb can be modeled as a circle with radius r and center (a,b). And thus a solar limb can be described with the parametric equations:

$$\begin{cases} x = a + r\cos\theta \\ y = b + r\sin\theta \end{cases} \tag{1.3}$$

When the angle θ steps through the 360 degree range, the points (x, y) trace the solar limb. If the radius r of a solar limb is known, its center location (a,b) can be figured out by constructing a Hough accumulation matrix. At first, a two-dimensional (2D) Hough accumulation matrix, which is of the same dimension as the digital image under consideration, is initialized to be all zeros. Edge points are figured out from the given image. For each edge point located at (*x,y*), the value of the corresponding elements of the accumulation matrix is increased by one, where the corresponding elements are on the perimeter of a circle, whose center location is (*x,y*) and radius is r. Finally, the element with the greatest value in the Hough accumulation matrix is found out, whose location is the center location of the circle. The processing is illustrated in Figure 1.5.



(a) (b)

Figure 1.5 Illustration of the Hough circle identification.

In reality, the radius of the limb of a solar disk in each Hα image is unknown. Let the radius be in the range of $[R_L, R_R]$. In order to determine the radius accurately, each possible radius is enumerated. Therefore, it is needed to construct a three-dimensional (3D) Hough accumulation matrix, in which each channel corresponds to an enumerated radius. The channel which contains the element of the greatest magnitude of the 3D Hough accumulation matrix identifies the radius r, and the location of the element identifies the center location (a,b). In Figure 1.6, the left panel displays a circle, and the right panel displays the channel of the accumulation matrix, containing the element of the highest magnitude, as a 3-dimensional mesh surface.

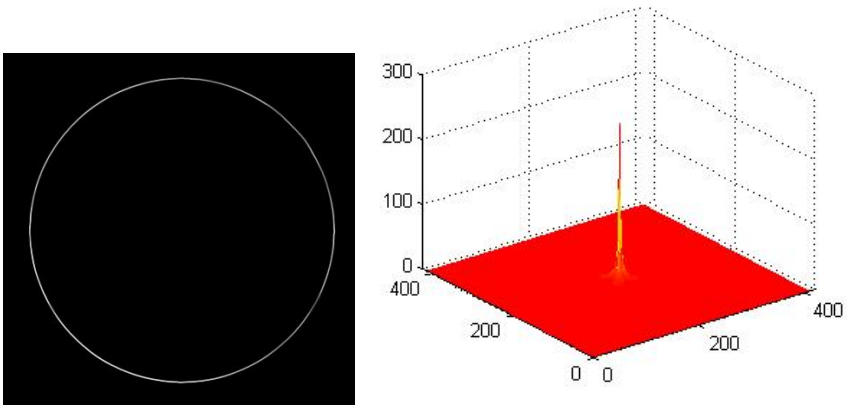


Figure 1.6 Illustration of the accumulation matrix of Hough circle transform.

1.5 Machine Learning

Machine learning [33] is a scientific discipline focused on the design and implementation of computer algorithms that allow computer systems to evolve based on data. A machine learning algorithm analyzes data to capture the characteristics of the data and thus it can make intelligent decisions based on the data [34].

Machine learning has been applied to various fields, including speech recognition [35, 36], natural language processing [37, 38], computer vision [39], and robotics [40]. Machine learning can be categorized into two categories: supervised learning and unsupervised learning [41]. In supervised learning, a learning model is established to map inputs to desired outputs. For example, support vector machine approximates a function mapping an input into a class by looking at input-output relations of the samples in a data set. Whereas unsupervised learning is to model a set of data, like clustering [42].

A supervised learning task usually involves separating data into training and testing sets. Each sample in the training set contains one "target value" (i.e. the class labels) and "several attributes" (i.e. the features, predictive parameters or observed variables) [43]. In this study, two supervised machine learning techniques are adopted for flare forecasting, namely logistic regression and support vector machine (SVM). The goal of logistic regression is to produce a model (based on the training data) which predicts the probabilities of target values of a testing data given only the testing data attributes, whereas the goal of SVM is to produce a model which predicts the target values of a testing data.

1.5.1 Logistic Regression

The logistic regression [44] is a machine learning technique to model the posterior probabilities of K classes via linear functions in input data, while at the same time

ensuring that the sum of the K posterior probabilities equals one and that each of the K posterior probabilities remain in $[0,1]$. The logistic regression model is presented in terms of $K-1$ log-odds, which has the following form [45]:

$$\begin{aligned}
\log\frac{\Pr(G=1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})} &= \beta_{10}+\boldsymbol{\beta}_1^T\mathbf{x}\\
\log\frac{\Pr(G=2\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})} &= \beta_{20}+\boldsymbol{\beta}_2^T\mathbf{x}\\
&\vdots\\
\log\frac{\Pr(G=K-1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})} &= \beta_{(K-1)0}+\boldsymbol{\beta}_{K-1}^T\mathbf{x}
\end{aligned} \tag{1.4}$$

The equations above can be transformed into the following equations:

$$\begin{aligned}
\frac{\Pr(G=1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})} &= \exp(\beta_{10}+\boldsymbol{\beta}_1^T\mathbf{x})\\
\frac{\Pr(G=2\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})} &= \exp(\beta_{20}+\boldsymbol{\beta}_2^T\mathbf{x})\\
&\vdots\\
\frac{\Pr(G=K-1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})} &= \exp(\beta_{(K-1)0}+\boldsymbol{\beta}_{K-1}^T\mathbf{x})
\end{aligned} \tag{1.5}$$

By adding the above equations, it can be derived that:

$$\begin{aligned}
&\frac{\Pr(G=1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})}+\frac{\Pr(G=2\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})}+\cdots\frac{\Pr(G=K-1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})}\\
&=\exp(\beta_{10}+\boldsymbol{\beta}_1^T\mathbf{x})+\exp(\beta_{20}+\boldsymbol{\beta}_2^T\mathbf{x})+\cdots+\exp(\beta_{(K-1)0}+\boldsymbol{\beta}_{K-1}^T\mathbf{x})
\end{aligned} \tag{1.6}$$

The equation above is equivalent to:

$$\begin{aligned}&\frac{\Pr(G=1\mid X=\mathbf{x})+\Pr(G=2\mid X=\mathbf{x})+\cdots+\Pr(G=K-1\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})}\\&=\exp(\beta_{10}+\boldsymbol{\beta}_1^T\mathbf{x})+\exp(\beta_{20}+\boldsymbol{\beta}_2^T\mathbf{x})+\cdots+\exp(\beta_{(K-1)0}+\boldsymbol{\beta}_{K-1}^T\mathbf{x})\end{aligned} \tag{1.7}$$

Under the assumption, the following condition must hold:

$$\Pr(G=1\mid X=\mathbf{x})+\Pr(G=2\mid X=\mathbf{x})+\cdots+\Pr(G=K\mid X=\mathbf{x})\equiv 1 \tag{1.8}$$

Equation 1.9 can be derived:

$$\frac{1-\Pr(G=K\mid X=\mathbf{x})}{\Pr(G=K\mid X=\mathbf{x})}=\exp(\beta_{10}+\boldsymbol{\beta}_1^T\mathbf{x})+\exp(\beta_{20}+\boldsymbol{\beta}_2^T\mathbf{x})+\cdots+\exp(\beta_{(K-1)0}+\boldsymbol{\beta}_{K-1}^T\mathbf{x}) \tag{1.9}$$

And then, the following equation can be calculated:

$$\begin{aligned}\Pr(G=K\mid X=\mathbf{x})&=\frac{1}{1+\exp(\beta_{10}+\boldsymbol{\beta}_1^T\mathbf{x})+\exp(\beta_{20}+\boldsymbol{\beta}_2^T\mathbf{x})+\cdots+\exp(\beta_{(K-1)0}+\boldsymbol{\beta}_{K-1}^T\mathbf{x})}\\&=\frac{1}{1+\sum_{i=1}^{K-1}\exp(\beta_{i0}+\boldsymbol{\beta}_i^T\mathbf{x})}\end{aligned} \tag{1.10}$$

In summary,

$$\Pr(G=k \mid X=\mathbf{x})=\begin{cases}\dfrac{\exp(\beta_{k0}+\beta_k^T\mathbf{x})}{1+\sum_{i=1}^{K-1}\exp(\beta_{i0}+\boldsymbol{\beta}_i^T\mathbf{x})} & k=1,2,\cdots,K-1 \\ \dfrac{1}{1+\sum_{i=1}^{K-1}\exp(\beta_{i0}+\boldsymbol{\beta}_i^T\mathbf{x})} & k=K\end{cases} \tag{1.11}$$

Before logistic regression can be used, it has firstly to be trained using training data by maximum likelihood [44], using the conditional likelihood of G given X . The log-likelihood for N data samples $(\mathbf{x}_i, g_i), i=1,2,\cdots,n$ is

$$l(\theta)=\sum_{i=1}^{N}\log\Pr(G=g_i \mid X=\mathbf{x}_i;\theta) \tag{1.12}$$

In two-class case ($K=2$), let $y_i=0$ when $g_i=1$ and $y_i=1$ when $g_i=2$, the log-likelihood above can be written as follows:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N}\{y_i\log\Pr(G=1 \mid X=\mathbf{x}_i;\theta)+(1-y_i)\log\Pr(G=2 \mid X=\mathbf{x}_i;\theta)\} \\ &= \sum_{i=1}^{N}\{y_i\log\Pr(G=1 \mid X=\mathbf{x}_i;\theta)+(1-y_i)\log(1-\Pr(G=1 \mid X=\mathbf{x}_i;\theta))\} \\ &= \sum_{i=1}^{N}\left\{y_i\theta^T\mathbf{x}_i-\log(1+e^{\theta^T\mathbf{x}_i})\right\} \end{aligned} \tag{1.13}$$

where $\theta=\{\beta_{10},\beta_1\}$, and it is assumed that the inputs $\mathbf{x}_i$ include a constant term 1 to accommodate the intercept.

To maximize the log-likelihood, set the derivative of the equation to zero as follows:

$$\frac{\partial l(\theta)}{\partial \theta} = \sum_{i=1}^{N} \mathbf{x}_i \left(y_i - \Pr(G = 1 \mid X = \mathbf{x}_i; \theta) \right) = 0 \tag{1.14}$$

Newton-Raphson algorithm can be used to solve the equation above. It can be derived that [45]:

$$\begin{cases} \theta^{new} = \theta^{old} - \left(\dfrac{\partial^2 l(\theta^{old})}{\partial \theta^{old} \partial (\theta^{old})^T} \right)^{-1} \dfrac{\partial l(\theta^{old})}{\partial \theta^{old}} \\ \dfrac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} = -\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T \Pr(G = 1 \mid X = \mathbf{x}_i; \theta) \left(1 - \Pr(G = 1 \mid X = \mathbf{x}_i; \theta) \right) \end{cases} \tag{1.15}$$

It seems that $\theta = 0$ is a good starting value for the iteration procedure. Typically the iteration procedure converges because the log-likelihood is concave [44, 46].

1.5.2 Support Vector Machine

Given a training set of attributes-label pairs $(\mathbf{x}_i, y_i)$, $i = 1, 2, ..., l$ where $\mathbf{x}_i \in R^n$ and $y_i \in \{1, -1\}$, a support vector machine (SVM) model is expressed as the following optimization problem [47, 48]:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$
$$subject\ to \begin{cases} y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i \\ \xi_i \geq 0 \end{cases} \tag{1.16}$$

where the training vectors $\mathbf{x}_i$ are mapped into a higher dimensional space by the function ϕ. The training of SVM model is to find a linear separating hyperplane with the maximal margin in this mapped higher dimensional space. $C > 0$ is a penalty parameter of the error term. The solution to the optimization problem above is given by the saddle point of the Lagrangian [49]:

$$\Phi(\mathbf{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\beta}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\left(y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i)+b\right)-1+\xi_i\right) - \sum_{j=1}^{l}\beta_i\xi_i \tag{1.17}$$

where $\boldsymbol{\alpha},\boldsymbol{\beta}$ are the Lagrange multipliers. The Lagrangian has to be maximized with respect to $\boldsymbol{\alpha},\boldsymbol{\beta}$ and minimized with respect to $\mathbf{w},b,\boldsymbol{\xi}$. The classical Lagrangian duality enables the primal problem above to be transformed to its dual problem as following [50]:

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left(\min_{\mathbf{w},b,\boldsymbol{\xi}}\left(\Phi(\mathbf{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\beta})\right)\right) \tag{1.18}$$

The minimum with respect to $\mathbf{w},b,\boldsymbol{\xi}$ of the Lagrangian Φ is given by the following equations:

$$\begin{aligned}
\frac{\partial \Phi}{\partial b} = 0 \quad &\Rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \\
\frac{\partial \Phi}{\partial \mathbf{w}} = 0 \quad &\Rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i) \\
\frac{\partial \Phi}{\partial \boldsymbol{\xi}} = 0 \quad &\Rightarrow \quad \alpha_i + \beta_i = C
\end{aligned} \tag{1.19}$$

From the above three equations, the dual problem is as follows [50]:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} -\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle - \sum_{k=1}^{l} \alpha_k \\
&subject\ to \begin{cases} 0 \le \alpha_i \le C, i = 1, 2, \cdots, l \\ \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases}
\end{aligned} \tag{1.20}$$

The Lagrange multipliers $\boldsymbol{\alpha}$ can be calculated by solving the above equation, and a SVM prediction model is given by [51]:

$$\begin{aligned}
&f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i \in SVs} \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + b \right) \\
&\text{where } b = -\frac{1}{2} \sum_{i=1}^{l} \alpha_i y_i \left(\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{r1}) \rangle + \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{r2}) \rangle \right)
\end{aligned} \tag{1.21}$$

Using the theory of kernel method [52], the mapping function ϕ does not need to be explicit. A kernel function $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)\phi(\mathbf{x}_j)$ can be used throughout the equations above. The most widely known kernel functions include the followings [53]:

1. Linear kernel function: $K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i^T \mathbf{x}_j$

2. Polynomial kernel function: $K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\gamma \mathbf{x}_i^T \mathbf{x}_j + r\right)^d, \gamma > 0$

3. Radial basis kernel function (RBF): $K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \gamma > 0$

4. Sigmoid kernel function: $K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \tanh\left(\gamma \mathbf{x}_i^T \mathbf{x}_j + r\right)$

CHAPTER 2

AUTOMATIC SOLAR FILAMENT SEGMENTATION AND CHARACTERIZATION

2.1 Introduction

In hydrogen alpha (Hα) full-disk solar images, solar filaments appear as elongated dark threads on the brighter solar disk as shown in Figure 2.1. Solar filaments (also called prominence when it appears at the solar limb) are clouds of relatively cool and dense gas suspended above the solar photosphere, generally along a magnetic neutral line [3]. Researchers are exploring the close relationship between erupting filaments and coronal mass ejections by studying the evolution of solar filaments [3, 6-8]. The Space Weather Research Lab (SWRL) at New Jersey Institute of Technology is currently maintaining a global high-resolution Hα network [5] which aims at maintaining a public accessible database containing all Hα images from different solar observatories around the world.

The geographically distributed observatories can perform 24 hours continuous observation to eliminate the limitation that one observatory can only observe about eight hours a day from sunrise to sunset. Since each of these observatories can produce hundreds of Hα full-disk solar images per day, it is a time-consuming and challenging task for observers to manually mark and measure features on the Sun, such as filaments. Besides, there is no complete and accurate solar filaments catalog available up to now, which is of vital importance for researchers on solar physics and space weather.

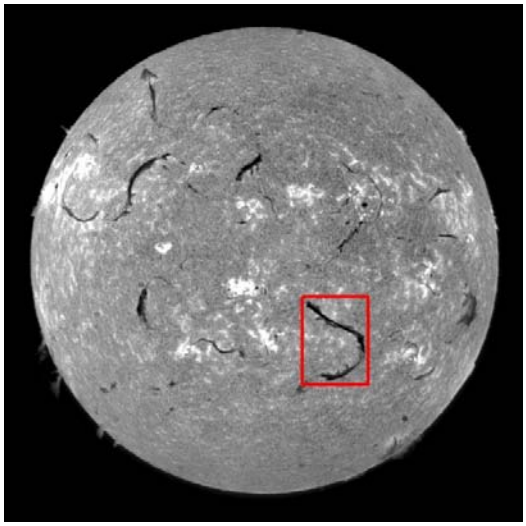


Figure 2.1 An Hα full-disk solar image (Courtesy of BBSO). One of the solar filaments is enclosed in a rectangle box.

Several automatic methods of solar filament segmentation have been proposed. Gao *et al*. [9] utilized global thresholding and region growing to segment filaments. Shih and Kowalski [10] developed both global and local thresholding combined with mathematical morphology to segment solar filaments. Qu *et al*. [12] developed an adaptive thresholding based on edge detection to detect solar filaments. Bernasconi *et al*. [11] used normalization and global thresholding to segment solar filaments. However, all of the methods above are designed and tested only on the solar images generated by the Big Bear Solar Observatory (BBSO) in California, and thus they may not able to work well on the solar images obtained by other solar observatories for the two reasons below.

First, the solar images produced by different observatories may have different properties, such as dynamic range, resolution, and luminance. A method which works fine with solar images produced by one observatory may not be suitable for those produced by other observatories. For example, Bernasconi *et al*. [11] selected the value −600 as the filament detection threshold for the solar images produced by BBSO, but this is definitely

not a good threshold for the solar images produced by Kanzelhöhe Solar Observatory (KANZ) in Austria, as illustrated in Figure 2.2.

Second, to calculate the longitude and latitude of the centroid of a solar filament, firstly it is required to know the center and radius of the solar disk in each image. The current methods use the center location and radius from the file header associated with each solar image. Unfortunately, it is found out that the center location and radius provided in the file header are not always very accurate. Furthermore, the format and description of the file header used by different observatories are different. For example, to describe the horizontal coordinate of the center location, the BBSO uses "CENX" but the KANZ uses "CRPIX1" in one image and use "CENTER_X" in another image. Sometimes, some images come with no such information at all. For example, one solar image produced by the Yunnan Astronomical Observatory (YNAO) in China provides no such information about the center location and radius of solar disk.

To tackle the aforementioned challenges, an adaptive segmentation method is put forward for solar filament segmentation and a cascading Hough circle detector for solar disk's center location and radius identification. The rest of the chapter is organized as follows. The procedure for solar disk's center location and radius identification is presented in Section 2. The procedure of solar filament segmentation is presented in Section 3. The characterization of solar filaments is discussed in Section 4. Section 5 illustrates the experimental results. Discussion and conclusion are included in the last section.

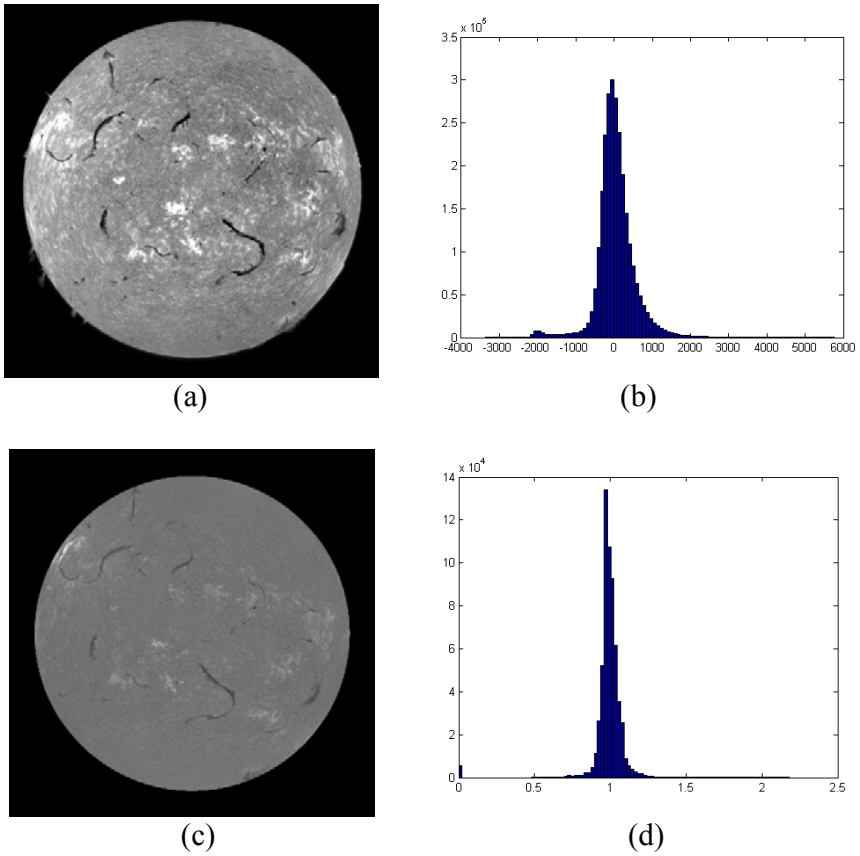


Figure 2.2 Solar images and their corresponding histograms (excluding background). (a) An Hα image taken by BBSO on Feb. 9th, 2002, (b) histogram of the solar disk in (a), (c) an Hα image taken by KANZ on Feb. 9th, 2002, (d) histogram of the solar disk in (c).

2.2 Identification of Center Location and Radius

The radii and the center locations of solar disks in Hα images vary from one image to another. The variation of the distance between the Earth and the Sun causes the variance of the radii. A slight error in telescope tracking contributes to the variance of center locations of solar disks.

It is a difficult task to accurately determine the center coordinates and radii of solar disks. The method designed by Denker *et al*. [54] has been adopted by BBSO in publishing Hα solar images. The method is easy to implement; however, it is vulnerable to noise on

the limb, especially when there are prominences extended above the limb in the four rectangular regions. In this section, a new method is presented which uses image smoothing, edge detection [55], and Hough transform [32, 56]. The workflow of the proposed solar radius and center identification algorithm is illustrated in Figure 2.3. The algorithm is composed of two major stages as described below.

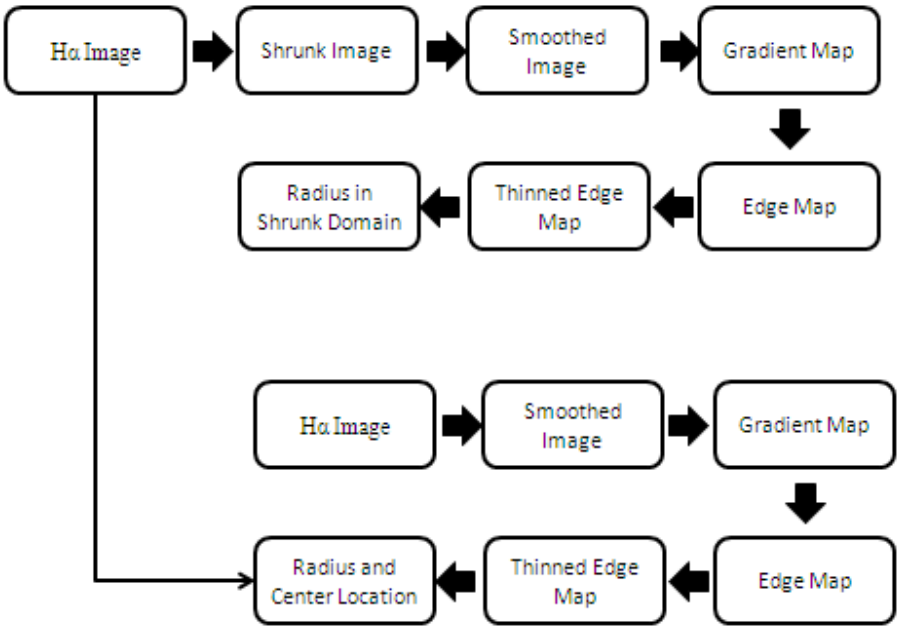


Figure 2.3 Workflow of solar radius and center location identification.

In stage one, first a given image is shrunk to $\frac{1}{k}$ of its original size. Second, the median filter is used to smooth the shrunk image. Third, an edge operator is applied to the smoothed image to obtain a gradient map. Fourth, the pixels in the gradient map whose intensities are greater than the median value plus three times of the standard deviation of pixel values on the filtered map are kept to obtain an edge map. Finally, the Hough circle detector is used to identify the solar radius R_{init} and center location (X_{init}, Y_{init}) of the shrunk

image, in which the radius search is performed within the range from $\frac{1}{4}$ to $\frac{1}{2}$ of the minimum of width and height of the given image. The underline assumption is that the radius of a solar disk is at least $\frac{1}{4}$ of the width or the height of the given image, whichever comes smaller, and the whole disk is enclosed in a given image, meaning that its radius is at most $\frac{1}{2}$ of the width or the height of the given image, whichever comes larger.

In stage two, first the given image is smoothed using the median filter. Second, an edge operator is applied to the smoothed image to produce a gradient map. Third, the pixels in the gradient map whose intensities are greater than the median value plus three times of the standard deviation of pixel values on the filtered map are kept to obtain an edge map. Finally, the Hough circle detector is used to identify the solar radius and center location, in which radius search is within the range of $[kR_{init}-k,\ kR_{init}+k]$.

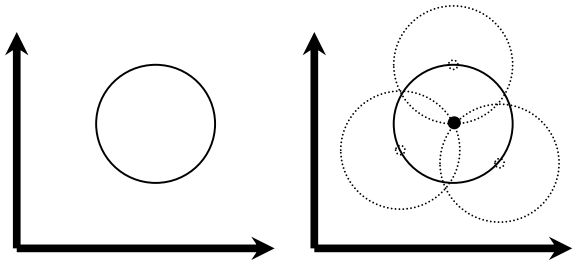
Hough circle detector works as follows: A circle with radius R and center location (a,b) can be described by the parametric equations:

$$\begin{cases} x = a + R\cos\theta \\ y = b + R\sin\theta \end{cases} \tag{2.1}$$

When the angle θ steps from zero to 360º, the points (x,y) form the perimeter of a circle. A given edge map I_{edge} contains the points (x,y) located on the solid circle in Figure 2.4 (a). Each of these points corresponds to the center of a circle which is illustrated as a dotted circle in Figure 2.4 (b). The location where dotted circles pass the most frequently

(marked as a solid black dot) or the location of the element whose value is greatest in the Hough accumulation matrix is the center location of the circle under investigation.

The whole process of Hough circle detector can be carried out by first constructing a two-dimensional (2D) Hough accumulation matrix, as initialized to be all zeros. For each foreground point in the edge map (where $I_{edge}(x,y) \equiv 1$), the value of the corresponding elements of the accumulation matrix is increased by one, where the elements are on the perimeter of a circle whose center location is (*x,y*) and radius is *R*. Finally, the element with the greatest value in the Hough accumulation matrix is found out, whose location is the center location of the circle.



(a) (b)

Figure 2.4 Illustration of the Hough circle identification.

The radius of the solar disk in each Hα image is a variable. Let the radius be in the range of $[R_L, R_R]$. In order to determine the radius accurately, each possible radius is enumerated. Therefore, it is needed to construct a three-dimensional (3D) Hough accumulation matrix, each of its channels corresponds to an enumerated radius. The channel which contains the element of the greatest magnitude of the 3D Hough accumulation matrix identifies the radius *R*, and the location of the element identifies the center location (*x,y*).

However, due to the following two constraints, direct Hough transform implementation is extremely inefficient in solar limb identification:

1. Memory limitation: The resolution of a Hα image is about 2000 by 2000, and the radii of solar disks are in the range from 500 to 1000 (the radius of a solar disk can be as small as 409, as one obtained by Catania Astrophysical Observatory (CAO), and also can be as large as 918, as one obtained by BBSO). Therefore, a 3D Hough accumulation matrix of 2000 rows, 2000 columns, and 1000-500+1≡501 channels is needed. If one 32-bit (i.e. 4 bytes) integer is used to represent each element of the accumulation matrix, the total memory consumption is $2000 \times 2000 \times 501 \div 1024 \div 1024 \div 1024 \times 4 \approx 7.4$ gigabytes, which are too big for most mainstream computers.
2. Computation limitation: Hough transform is applied on bi-level images (edge map). The computational complexity of Hough transform is proportional to the number of "1"s in the thinned edge map (Note: "0" as background). Supposing that one twentieth of the original image pixels are detected as edge points, the number of edge points detected would be 2032×2032/20 ≈ 200000. Each of these edge points will go through one iteration in the Hough accumulation matrix building up, which consumes a lot of time.

To mitigate the two limitations, a cascading two-stage approach is developed. In the first stage, a given image (in which the radius of solar disks is within $[R_L, R_R]$) is shrunk to $\frac{1}{k^2}$ of its original size. Let *I* be the original image of *M* rows by *N* columns and

its shrunk image be I_r of $\frac{M}{k}$ rows by $\frac{N}{k}$ columns. After median filtering and Roberts edge operator, Hough transform is used to identify the solar limb in the shrunk image whose radius is within $\left[\frac{R_L}{k}, \frac{R_R}{k}\right]$. Let the radius of the shrunk image detected by the Hough transform be R_T, so the radius R of the original image should lie within $\left[kR_T - k, kR_T + k\right]$.

In the second stage, the median filtering, Roberts edge detection, edge thinning, and Hough transform are performed on the original image I. However, the search of the radius is performed within $\left[kR_T - k, kR_T + k\right]$, that saves computational time significantly.

For example, the resolution of an Hα full-disk solar image produced by BBSO is 2032 by 2032, in which the radius of solar limb is within [500, 1000]. As aforementioned, the direct Hough circle identification needs to construct a Hough matrix consuming around seven gigabytes of memory. Using the two-stage Hough circle identification approach, the memory consumption is as follows.

In the first stage, the image is supposed to be shrunk to $\frac{1}{25}$ of its original size. The radius in the shrunk image would be within [100, 200]. It is needed to construct a 3D Hough accumulation matrix of size $2032/5 \equiv 407$ rows, $2032/5 \equiv 407$ columns, and $200 - 100 + 1 \equiv 101$ channels. The memory consumption of the matrix is $407 \times 407 \times 101 \div 1024 \div 1024 \times 4 \approx 63$ megabytes.

In the second stage, suppose the identified radius from the first stage is 180. The radius of the solar limb in the original image should be within $\left[180 \times 5 - 5, 180 \times 5 + 5\right] \equiv \left[895, 905\right]$. It is needed to construct a 3D Hough accumulation

matrix of size 2032 rows, 2032 columns, and $905-895+1\equiv 11$ channels, which consumes $2032\times 2032\times 11\div 1024\div 1024\times 4\approx 173$ megabytes of memory.

Figure 2.5 and 2.6 illustrate an example of the first and second stages of solar radius and center identification procedure, respectively.

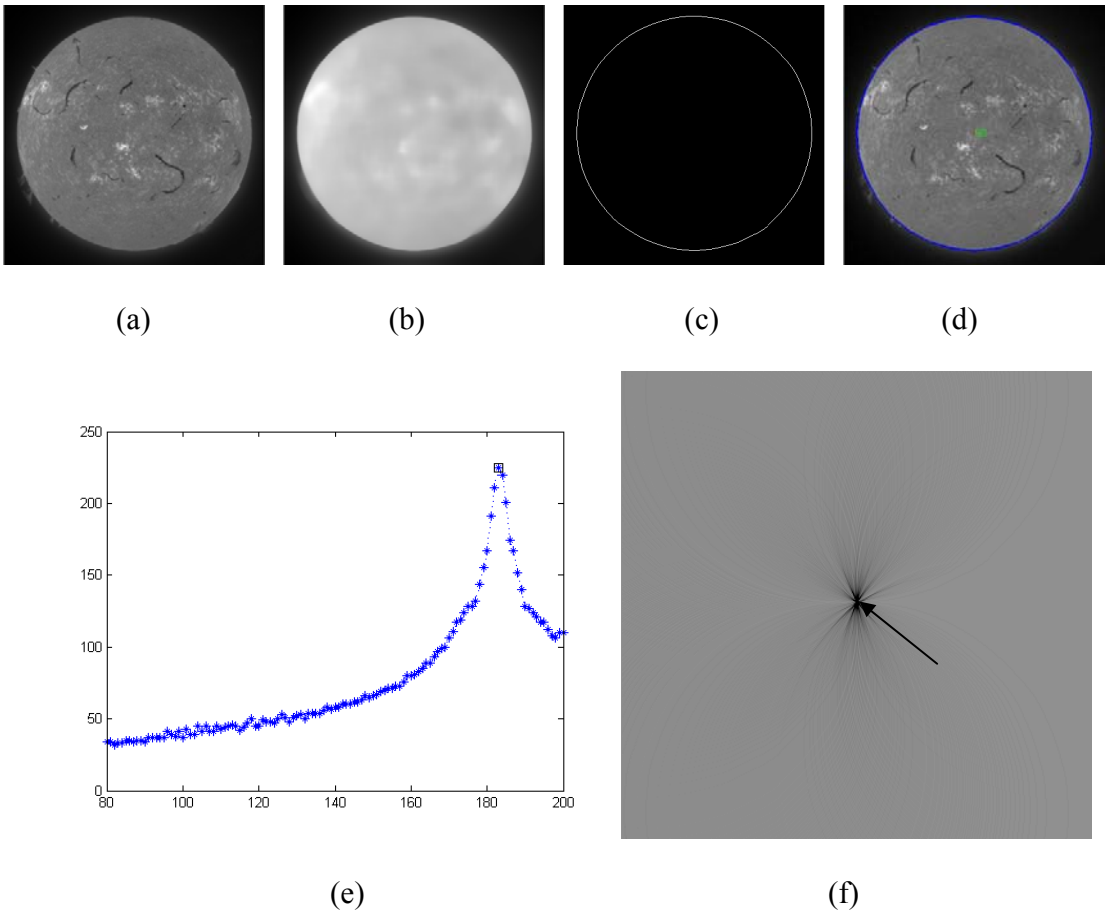
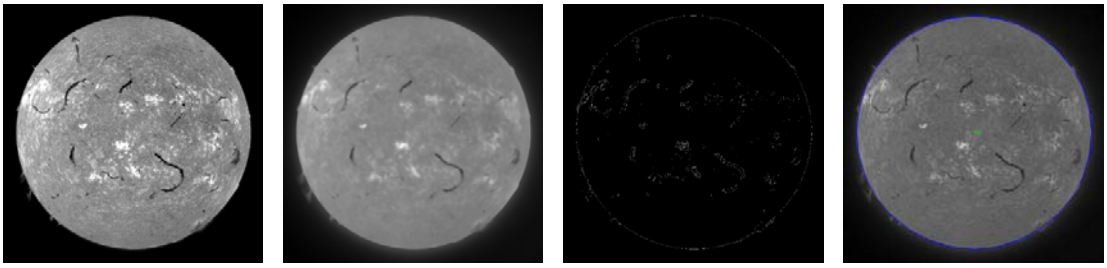
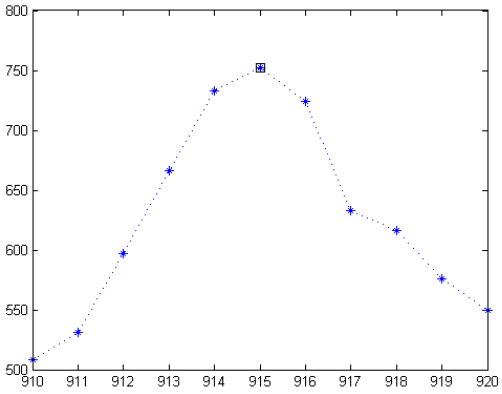
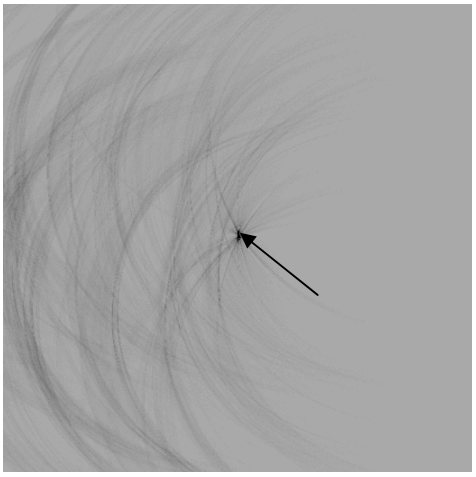


Figure 2.5 An example of the first stage of solar radius and center identification. (a) Shrunk image, (b) shrunk image after median filtering, (c) thinned edge map, (d) identified solar limb over plotted on the solar image as a blue circle, (e) a dotted curve on which each star mark corresponds to the maximum value of each channel of Hough accumulation matrix, (f) the 183rd channel of Hough accumulation matrix shown as an image, in which the darkest point (pointed to by an arrow) on the center area illustrates the center location of the identified solar disk.



(a) (b) (c) (d)





(e) (f)

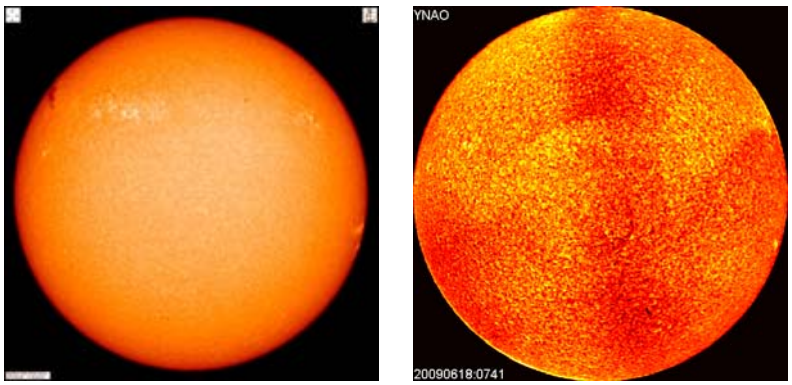
Figure 2.6 An example of the second stage of solar limb identification. (a) Original image, (b) the image after median filtering, (c) thinned edge map, (d) identified solar limb over plotted on the solar image as a blue circle, (e) a dotted curve on which each star mark corresponds to the maximum value of each channel of Hough accumulation matrix, (f) The 915th channel of Hough accumulation matrix shown as an image, in which the darkest point (pointed to by an arrow) on the center area illustrates the center location of the identified solar disk.

2.3 Segmentation of Solar Filament

2.3.1 Unbalanced Luminance Correction

When the Sun is observed in hydrogen alpha line, it is noticed that the brightness of the disk gradually decreases from its center to its limb (shown as Figure 2.7(a)), which is called

limb darkening [57]. After limb darkening removal [54], the background luminance of the solar disks in Hα images is still non-uniform; some location is brighter than other location (illustrated in Figure 2.7(b)). This may result from (1) clouds in the atmosphere of the Earth, which blocks the sunlight at some location, (2) the dusts on the telescope, (3) the dusts on the film, or (4) the electronic noise brought in during the film digitization procedure. Note that historical Hα images were recorded in traditional 35mm film instead of digital camera nowadays.



(a) (b)

Figure 2.7 Illustration of unbalanced luminance. (a) An Hα image with limb darkening (Courtesy of BBSO), (b) unbalanced luminance on solar disk (Courtesy of YNAO).

Let $f(x,y)$ be an Hα full-disk solar image, which can be viewed as the combined effect of true luminance from the Sun $h(x,y)$ and noise luminance $g(x,y)$, whose relationship can be viewed as $f(x,y)=g(x,y)+h(x,y)$. The noise luminance $g(x,y)$ can be modeled as a polynomial function. If $g(x,y)$ is know, the true luminance from the Sun $h(x,y)$ can be obtained as $h(x,y)=f(x,y)-g(x,y)$.

$g(x,y)$ can be approximated using a polynomial function. The coefficients of the polynomial function can be figured out by minimizing the mean square difference between $f(x,y)$ and $g(x,y)$ as follows:

$$d(\boldsymbol{\alpha}) = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y) - g(x,y)\right)^2 \tag{2.2}$$

Here an example is illustrated by showing the computation of the coefficients of $g(x,y)$ using a third-degree polynomial function. That is:

$$g(x,y) = \alpha_0 + \alpha_1 x + \alpha_2 y + \alpha_3 x^2 + \alpha_4 y^2 + \alpha_5 xy + \alpha_6 x^3 + \alpha_7 y^3 + \alpha_8 x^2 y + \alpha_9 xy^2 \tag{2.3}$$

The mean square difference is:

$$\begin{aligned} d(\boldsymbol{\alpha}) &= \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y) - g(x,y)\right)^2 \\ &= \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y) - \left(\alpha_0 + \alpha_1 x + \alpha_2 y + \alpha_3 x^2 + \alpha_4 y^2 + \alpha_5 xy + \alpha_6 x^3 + \alpha_7 y^3 + \alpha_8 x^2 y + \alpha_9 xy^2\right)\right)^2 \end{aligned} \tag{2.4}$$

To calculate $\boldsymbol{\alpha}$, take partial differentiation of $d(\boldsymbol{\alpha})$ and set it to be 0. The matrix form $\mathbf{H}\boldsymbol{\alpha} = \mathbf{w}$ can be derived as follows:

$$\begin{bmatrix} \sum_{x=1}^{M}\sum_{y=1}^{N}1 & \sum_{x=1}^{M}\sum_{y=1}^{N}x & \sum_{x=1}^{M}\sum_{y=1}^{N}y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}y & \sum_{x=1}^{M}\sum_{y=1}^{N}xy & \sum_{x=1}^{M}\sum_{y=1}^{N}y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y & \sum_{x=1}^{M}\sum_{y=1}^{N}y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y & \sum_{x=1}^{M}\sum_{y=1}^{N}y^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^3 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^5 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^2 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^5 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^4 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}xy & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^3 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^5 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^6 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^5y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y^2 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^5 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}y^6 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^5 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^5y & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^3 \\ \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^4 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^4y^2 & \sum_{x=1}^{M}\sum_{y=1}^{N}xy^5 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^3y^3 & \sum_{x=1}^{M}\sum_{y=1}^{N}x^2y^4 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \\ \alpha_9 \end{bmatrix} = \begin{bmatrix} \sum_{x=1}^{M}\sum_{y=1}^{N}f(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}xf(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}yf(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x^2f(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}y^2f(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}xyf(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x^3f(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}y^3f(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}x^2yf(x,y) \\ \sum_{x=1}^{M}\sum_{y=1}^{N}xy^2f(x,y) \end{bmatrix} \tag{2.5}$$

Therefore, $\boldsymbol{\alpha}$ can be obtained by $\boldsymbol{\alpha} = \mathbf{H}^{-1}\mathbf{w}$. By taking the difference between $f(x,y)$ and $g(x,y)$, the luminance corrected image can be obtained.

2.3.2 Solar Filament Segmentation

Solar filaments differ in shape and luminance, which makes it difficult to segment them. Accordingly, an adaptive solar filament segmentation algorithm, which aims to be applicable to solar images produced by different solar observatories, is designed. The segmentation technique can be adapted to solar images with different dynamic range and statistical properties. The work flow of the proposed adaptive segmentation algorithm is illustrated in Figure 2.8 and explained as follows.

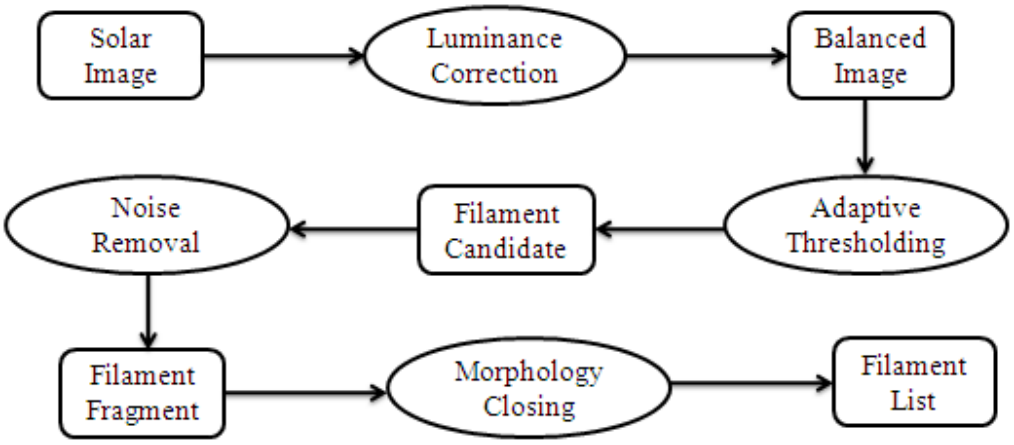


Figure 2.8 The work flow of solar filament segmentation algorithm.

First, for a given Hα image $f(x,y)$, calculate $g(x,y)=f(x,y)*h(x,y)$, which is the convolution of $f(x,y)$ with a high pass Laplacian filter $h(x,y)=\begin{bmatrix}1 & 1 & 1\\ 1 & -8 & 1\\ 1 & 1 & 1\end{bmatrix}$.

Second, calculate the median value V_{med} and standard deviation V_{std} of the set of pixels inside the solar disk in $f(x,y)$. Generate a series of thresholds which are composed of arithmetic progression between $(V_{med}-3V_{std})$ and V_{med}, as given by: $T_i=(V_{med}-3V_{std})+\frac{3V_{std}}{S}i$, $i=1,2,3,\cdots,S$, $S\in \mathrm{Z}$.

Third, by segmenting $f(x,y)$ using threshold T_i. Region $r_i(x,y)=\begin{cases}1 & f(x,y)<T_i\\ 0 & f(x,y)\geq T_i\end{cases}$ can be obtained, where $i=0,1,2,3...,S$, and then difference region $d_i(x,y)=r_{i+1}(x,y)-r_i(x,y)$ can be obtained, where $i=0,1,2,3...,S-1$.

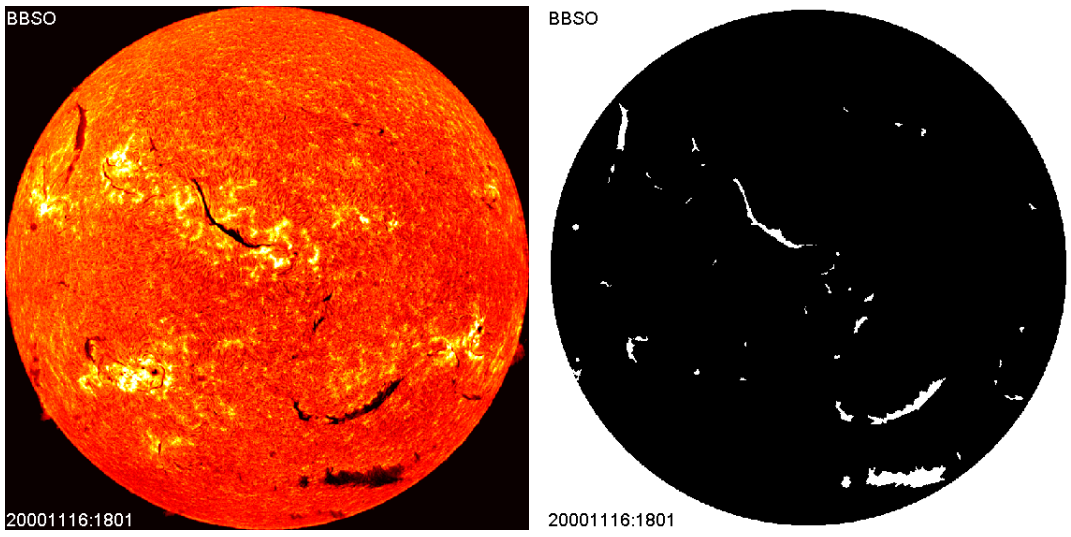
Fourth, calculate $J_i=\frac{\sum\sum d_i(x,y)\,g(x,y)}{\sum\sum d_i(x,y)}$, $i=0,1,2,3...,S-1$.

Fifth, among the series of $\{J_0, J_1, J_2, \cdots J_{S-1}\}$, search for the index k where J_k is maximum and segment $f(x,y)$ using threshold T_k to obtain the resulting filaments candidates map: $m(x,y) = \begin{cases} 1 & f(x,y) < T_k \\ 0 & f(x,y) \geq T_k \end{cases}$, where $m(x,y)$ which is a binary map, with '1' indicating object and '0' indicating background. Each 8-connected component is treated as a filament candidate.

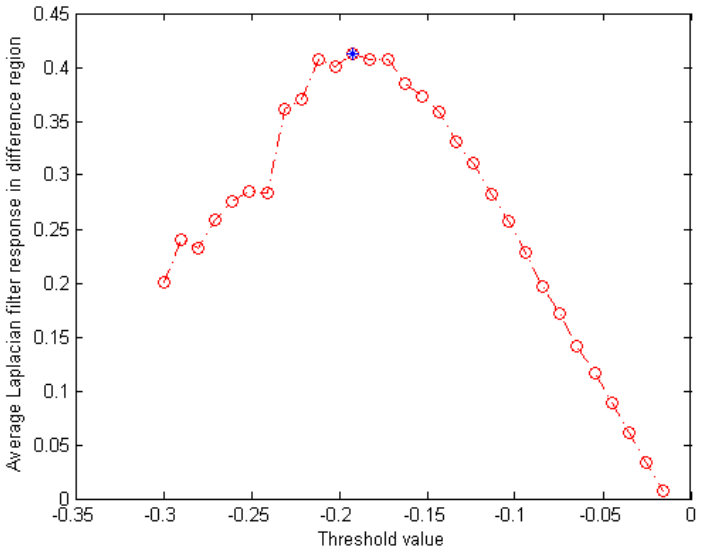
Sixth, remove those 8-connect components if their areas are less than β times the whole area of the solar disk. Let the result be $m'(x,y)$.

Finally, apply mathematical morphology closing [29] on filament map $m'(x,y)$ with a disk structuring element $n(x,y)$, whose radius is P, to connect broken filaments to get the final filaments map $m''(x,y)$.

Figure 2.9 shows an example of filament segmentation.



(a) (b)



(c)

Figure 2.9 Illustration of filament segmentation. (a) An Hα image taken on Nov. 16th of 2000 (courtesy of BBSO), (b) filaments found out by the proposed algorithm, (c) a curve on which the red circles, denoted as (T_i, J_i), are derived by the proposed algorithm in steps 4. The blue star on the peak of the curve denotes the best threshold value obtained by the proposed algorithm.

2.4 Characterization of Solar Filament

Four properties are used to describe each piece of solar filament, which are computed as follows:

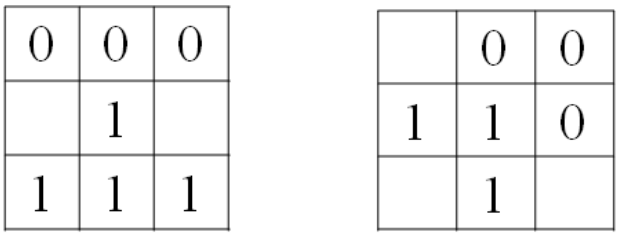
Area: The diameter of the Sun is about 1,392,000 kilometers. Let the radius of the solar disk in a given image be *R* pixels and the total number of pixel of a given filament be *N*. The area of the given filament in square kilometers can be computed by

$$N\left(\frac{1392000}{2R}\right)^2 \tag{2.6}$$

Location: Suppose that the centroid of a filament lies on (*x,y*) and the center of the solar disk lies on (x_c, y_c) using the origin on the upper-left corner of each solar image. Then convert the centroid location to longitude *lon* and latitude *lat* representation as follows [58]:

$$\begin{cases} lat = -\arcsin\left(\dfrac{y-c_y}{r}\right)\times\dfrac{180}{\pi} \\ lon = \arcsin\left(\dfrac{x-c_x}{\sqrt{r^2-\left(y-c_y\right)^2}}\right)\times\dfrac{180}{\pi} \end{cases} \tag{2.7}$$

Prior to measuring the length and slope of a filament, a further process is performed. First, fill the holes inside each filament using morphological reconstruction proposed in [59]. Second, obtain the skeleton of a filament by iterative mathematical morphology thinning [29]. At each iteration, the image is firstly thinned by the structure element in Figure 2.10 (a) and then by the structure element in Figure 2.10 (b), and followed by the remaining six 90^o rotations of the two structure elements. The process is repeated in a cyclic fashion until none of the thinning produces any further change. This procedure can produce a connected single-pixel width skeleton for each piece of filament.



(a) (b)

Figure 2.10 Structure elements for mathematical morphological thinning.

Length: Since the filament skeleton generally contains a lot of small branches or barbs, a robust method is designed based on graph theory to find out the main skeleton by removing smaller branches or barbs. First, create a graph (adjacent matrix representation) H for each filament. For each pixel on the skeleton, create a vertex (numbering) representing it. Then create an unweighted undirected graph H which contains all the vertices and connectivity of the vertices. If two pixels of the skeleton are 8-connected, create an edge connecting the two vertices corresponding to the two pixels.

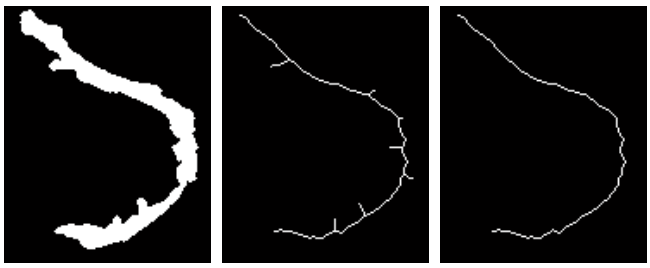
After transforming each filament skeleton into a graph, use a graph algorithm to find out the main skeleton. The main skeleton is defined as the longest acyclic path which connects two vertices. Two algorithms are designed to find out the main skeleton.

Algorithm 1: find out the all-pairs shortest path between any pairs of vertices using the Floyd-Warshall algorithm [60]. The path with maximum length is the main skeleton.

Algorithm 2: find out all the end vertices. An end vertex is defined as a vertex which has only one edge associated with it. Then use Dijkstra's single source shortest path algorithm [60] to search for the shortest path between each pair of these end vertices. The path with maximum length is the main skeleton.

The first algorithm is easy to implement, with time complexity of $O\left(V^3\right)$, where V is the number of vertices in a graph. The second algorithm is more efficient since its time complexity is $O\left(V^2E\right)$, where E is the number of end vertices and generally $E << V$.

Suppose that main skeleton contains Q vertices. The length of the corresponding filament in kilometers is $Q\left(\frac{1392000}{2R}\right)$, where the radius of the solar disk in a given image is R pixels. Figure 2.11 shows an example of main skeleton detection.



(a) (b) (c)

Figure 2.11 An illustration of main skeleton finding. (a) A piece of filament, (b) skeleton with barbs produced by mathematical morphology thinning, (c) the main skeleton.

At first sight, it seems unnecessary to derive the main skeleton. However, after removing the barbs, it is a trivial job to figure out the length of a piece of filament. The length of a piece of filament is just the total length of the path from one end to another end.

Slope: Slope is defined as the tangent of the angle between the horizontal line and the line connecting the two ends of the main skeleton. The angle is illustrated in Figure 2.12.

Figure 2.12 The tangent of angle β is used as slope of the main skeleton.

2.5 Experimental Results

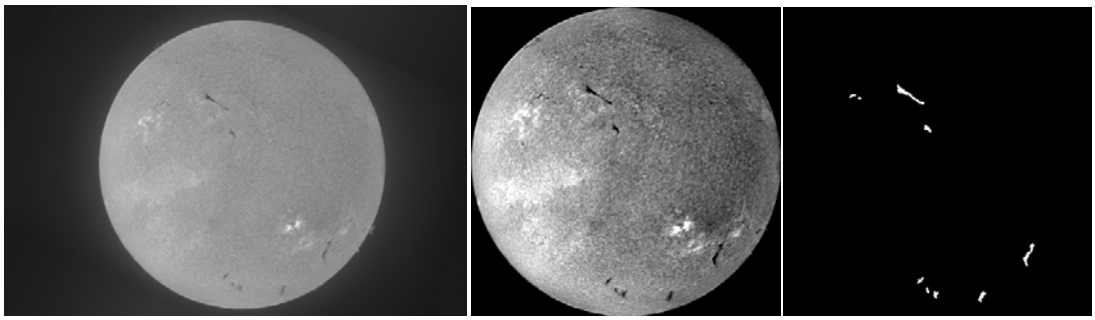
2.5.1 Dataset

To test the performance of the proposed methods, a dataset composed of 125 images generated by four different solar observatories is established, namely Big Bear Solar Observatory in California (BBSO), Kanzelhöhe Solar Observatory in Austria (KANZ), Catania Astrophysical Observatory in Italy (OACT), Yunnan Astronomical Observatory in China (YNAO). To be representative with respect to different time, the dataset is setup by choosing one image per month from January 2000 to May 2010. The dataset can be accessed at [61].

During the time period from January 2000 to May 2010, there are totally 125 months. Among the 125 months, Global H-alpha Network has shown that KANZ contributed images in 94 months, BBSO contributed images in 89 months, OACT contributed images in 78 month, and YNAO contributed images in 14 months. The ratio

between the four stations is $94:89:78:14$, which is approximately $1:0.95:0.83:0.15$. In our dataset, solar images in the dataset are chosen according to this ratio. Among the 125 images, 44 images are selected from KANZ, 40 from BBSO, 30 from OACT, and 11 from YNAO. The ratio between the number of images is $44:40:30:11 \approx 1:0.9:0.7:0.25$.

For each solar image, the solar radius and center location are manually identified, the image is cropped to contain only solar disk, and then all solar filaments presented on the solar disk are marked. This is for the comparison with automatically identified solar filaments. All the hand-marked filament maps and hand-cropped images can be accessed at [62]. Figure 2.13 shows a sample of manually cropped image and manually marked solar filaments.



(a) (b) (c)

Figure 2.13 Illustration of filament segmentation by hand. (a) An Hα solar image taken on March 20, 2010 by Yunnan Astronomical Observatory, (b) manually cropped image containing only solar disk, (c) manually marked filament.

2.5.2 Evaluation of Solar Radius and Center Location Identification

The success of the proposed circle detection largely depends on the quality of edge points detected. Good quality means that the detected edge points are located right on the boundary of solar limb and there is little noise presence. To measure the quality of the edge

points detected, a measurement Q_e is designed which is the ratio between the number of edge points located right on the edge boundary and the total number of edge points detected.

Assume that the dimension of a given hydrogen alpha full disk image is $M \times N$. Firstly, manually identify the radius *R* and center location (*X,Y*) of the containing solar disk. Then create a binary map with the same dimension as the image of $M \times N$, containing a circle centered at (*X,Y*) with radius *R*. Let the radius of one solar disk be *R*, and the horizontal and vertical location of its center be x_c and y_c, respectively. The circle map $f_c(x,y)$ is generated as follows:

$$f_c(x,y) = \begin{cases} 1 & \left\| \sqrt{(x-x_c)^2 + (y-y_c)^2} - R \right\| < 2 \\ 0 & \left\| \sqrt{(x-x_c)^2 + (y-y_c)^2} - R \right\| \geq 2 \end{cases} \tag{2.8}$$

Let the edge map $f_e(x,y)$ be the result of a given image after applying the proposed edge detection and edge thinning method. Let $f_e(x,y)$ be a binary map, where "1" means edge point and "0" means background. Then the quality measurement Q_e is computed as:

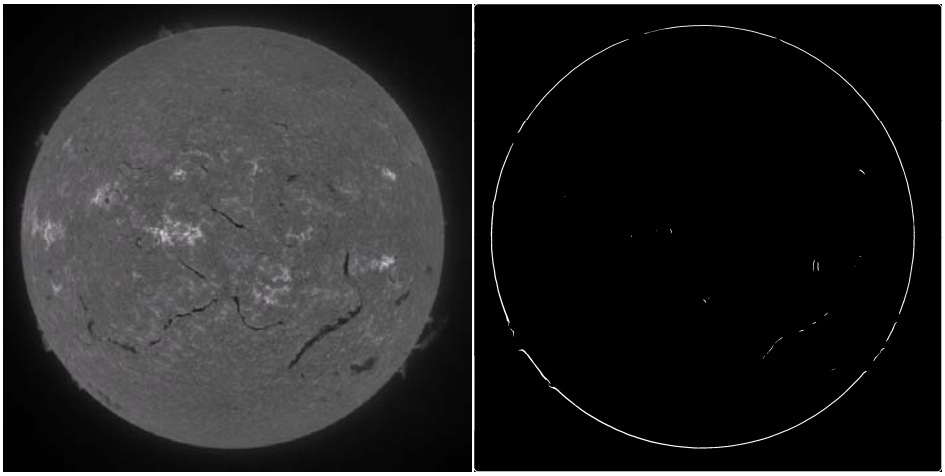
$$Q_e = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\left[f_e(x,y) \equiv f_c(x,y) \equiv 1 \right]}{\sum_{x=1}^{M}\sum_{y=1}^{N} f_e(x,y)} \tag{2.9}$$

Experimental results are illustrated in Table 2.1, containing average Q_e on the 125 images using eight different combinations for edge detection. The eight combinations are the permutation of one from the two smoothing filters (Gaussian filter [29] and median filter [63, 64]) and the other from the four edge operators (i.e., Roberts operator, Sobel operator, LoG operator, and Canny operator [65]). In the table, each row illustrates the quality measure of one combination with respect to the change of smoothing filter size. The following notations are used: "G" means Gaussian filter, "M" means median filter, "R" means Roberts edge operator, "S" means Sobel edge operator, "L" means Log edge operator, and "C" means Canny edge operator. Each column corresponds to a different filter size, and its deviation is chosen as a half of filter size for Gaussian filter.

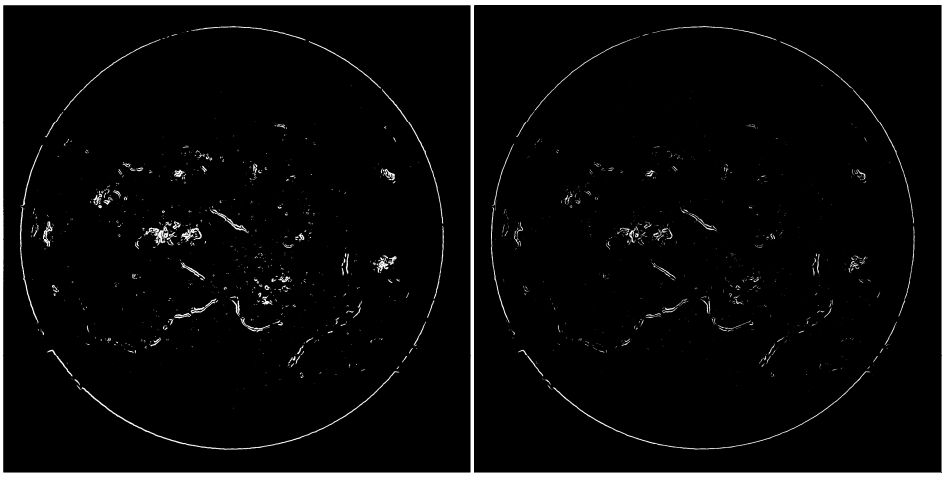
From experiments, the highest quality is 0.74, indicating that the median filter with size 35 combined with Roberts operator produces the best edge map for solar limb. Therefore, this combination is chosen in the proposed edge detection procedure. Figure 2.14 illustrates the results of three different edge detection combinations.

Table 2.1 Average Quality of Edge Points

	5	10	15	20	25	30	35	40
G/R	0.55	0.25	0.15	0.11	0.09	0.08	0.07	0.06
G/S	0.50	0.38	0.18	0.12	0.09	0.08	0.07	0.07
G/L	0.30	0.10	0.03	0.00	0.00	0.00	0.00	0.00
G/C	0.47	0.37	0.27	0.14	0.10	0.08	0.07	0.07
M/R	0.62	0.64	0.70	0.67	0.71	0.68	**0.74**	0.69
M/S	0.56	0.61	0.63	0.63	0.64	0.64	0.65	0.64
M/L	0.47	0.59	0.53	0.61	0.51	0.60	0.50	0.59
M/C	0.55	0.55	0.61	0.57	0.62	0.58	0.62	0.59



(a) (b)



(c) (d)

Figure 2.14 Edge detection results. (a) An Hα full disk image taken on Jan 22, 2001 at BBSO, (b) edge detection using Median filter (size 30) with Canny operator, (c) edge detection using Gaussian filter (size 5) and Canny operator, (d) edge detection using Median filter (size 10) with Roberts operator.

An IDL [66] program called find_limb.pro which can identify the center location and radius of the solar disk in an image has been chosen for comparison with the performance of the proposed method. The reason for choosing that program is that find_limb.pro is a program in the SolarSoftWare (SSW) library [67] within solar physics

community. Find_limb.pro uses Sobel edge operator to figure out the edge points, and then tries to fit a circle on the derived edge points.

Suppose the real center location of a solar disk is (x_1, y_1), the real radius is r_1, the center location and radius figured out from find_limb.pro or the proposed method are (x_2, y_2) and r_2, $err_1 = \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{r_1}$ is used to measure the error on center location identification, and $err_2 = \frac{|r_1 - r_2|}{r_1}$ is used to measure the error on radius identification.

For test results using Find_limb.pro, the mean value of err_1 is 0.0201, and standard deviation of err_1 is 0.1180; the mean value of err_2 is 0.0120, and standard deviation of err_2 is 0.0338.

For test results using our proposed method, the mean value of err_1 is 0.00044, and standard deviation of err_1 is 0.00099; the mean value of err_2 is 0.00014, and standard deviation of err_2 is 0.00053.

2.5.3 Solar Filament Segmentation Accuracy Measure

Figure 2.15 shows an example of solar filament segmentation and characterization. In Figure 2.15 (a), the radius and center location are identified, and then the image is cropped to produce Figure 2.15 (b). After applying the proposed segmentation method, a binary map in Figure 2.15 (c) is obtained, showing solar filaments which are thinned into Figure

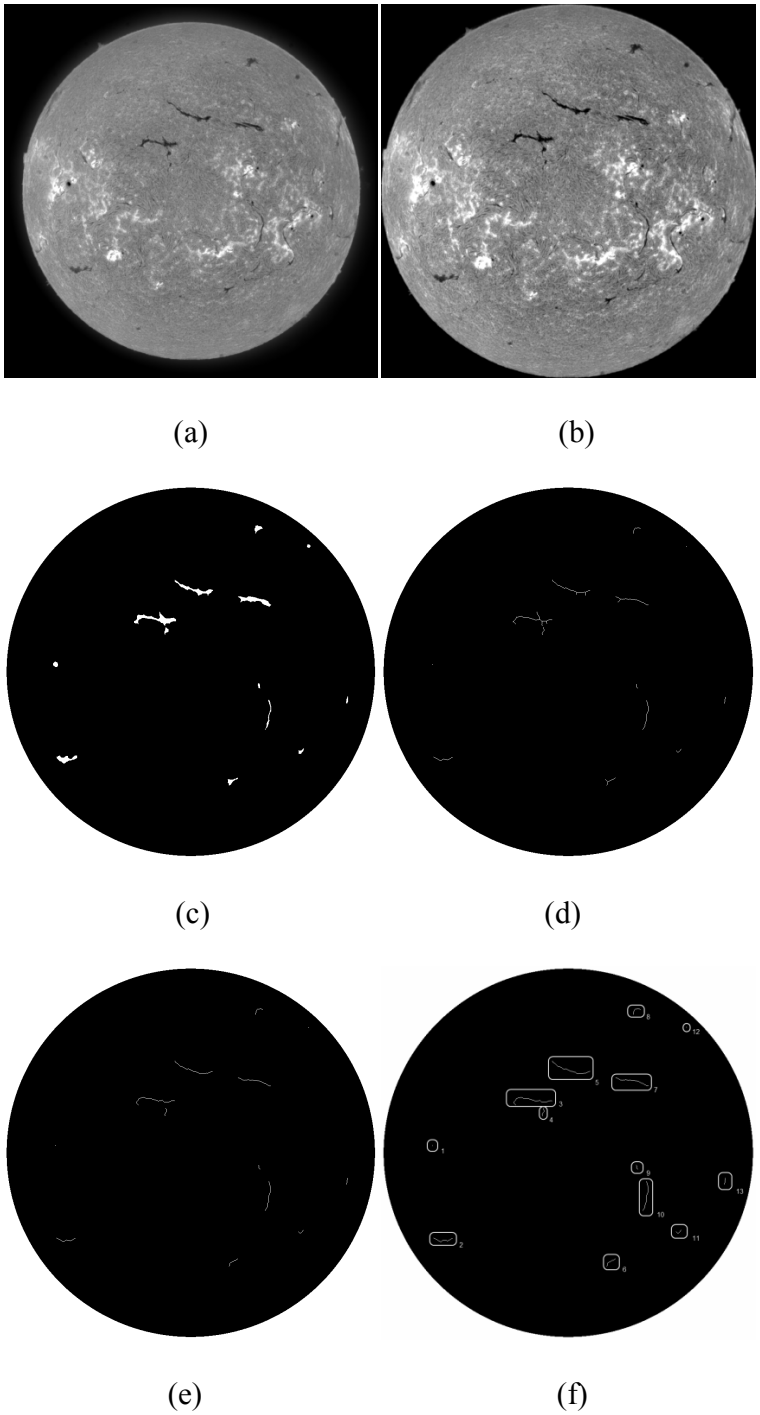


Figure 2.15 Illustration of solar filament segmentation. (a) An Hα image taken on January 20, 2002 by BBSO, (b) cropped to enclose only solar disk, (c) solar filaments segmented, (d) skeleton of solar filaments, (e) main skeletons of solar filaments, (f) indexed skeletons of solar filaments.

2.15 (d). Their main skeletons are identified as shown in Figure 2.15 (e). The characterization result is listed in Table 2.2, where the ID number of each filament corresponds to the number in Figure 2.15 (f) and the area is calculated for the corresponding filaments in Figure 2.15 (c).

Table 2.2 Characterized Solar Filaments

	Area, Mm^2	Longitude, degree (West positive)	Latitude, degree (North positive)	Length, Mm	Slope, tan β
1	284.67	-46.34	2.27	3.09	Infinite
2	1100.40	-49.03	-27.79	71.15	-0.02
3	2365.86	-11.10	16.35	157.76	0.11
4	234.43	-7.69	12.68	29.39	9.00
5	1550.13	1.58	26.06	143.84	-0.28
6	399.49	15.52	-35.83	41.76	0.86
7	1715.19	21.81	22.46	122.19	-0.26
8	454.51	33.91	50.01	40.21	0.71
9	119.61	21.46	-4.29	17.01	-2.25
10	447.34	25.41	-13.37	109.81	7.78
11	208.12	40.64	-25.24	21.65	0.33
12	131.57	58.38	42.23	3.09	0.00
13	131.57	57.82	-8.77	24.75	7.50

To evaluate the accuracy of the proposed filament segmentation algorithm, the filament maps generated by computer is compared with those manually generated. Two measures are used. The first measure is number-ratio, which is the ratio between the number of filaments marked by hand overlapping with those by the proposed method and the total number of filaments marked by hand. The number-ratio shows the percentage of the correctly identified number of solar filaments. The second measure is area-ratio, which is the ratio between the area of the algorithm-identified solar filaments and the area of the

hand-marked solar filaments. The area-ratio shows the percentage of the correctly identified area of solar filaments.

Figure 2.16 illustrates the two measures. Assume that the two red regions are the two filaments marked by human, the two blue regions are the two filaments obtained by the proposed algorithm, and the yellow region is the overlapping region of a filament by hand and by the proposed algorithm. Note that the yellow region is a subset of the red or blue region. The number-ratio is 0.5 in this case and the area-ratio is the ratio between the area of the yellow region and the total area of the red plus yellow regions.

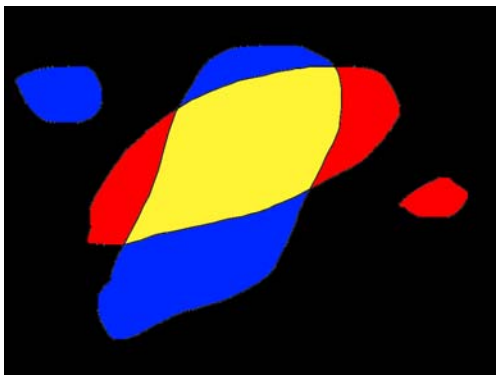
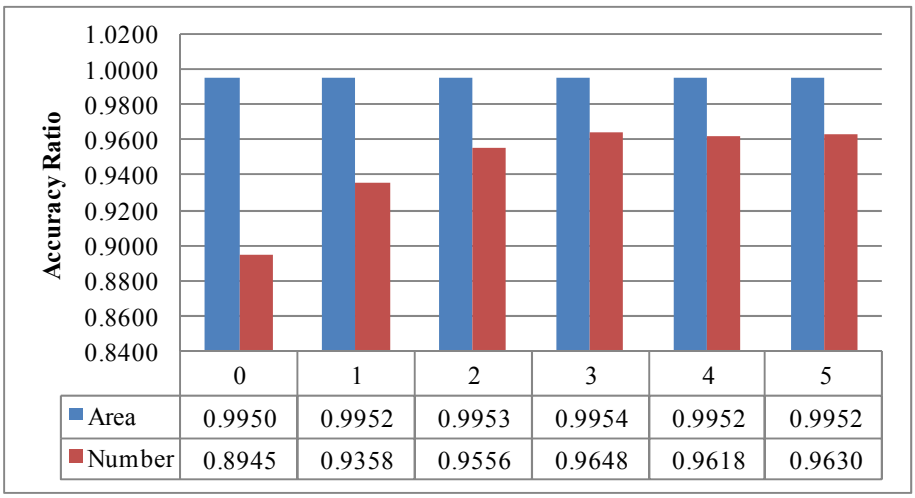


Figure 2.16 Illustration of filaments marked by hand (red color) and those obtained by the proposed algorithm (blue color). The yellow region is the overlapping region of a filament both by hand and by the proposed algorithm.

The proposed algorithm can be fine tuned by adjusting several parameters. By changing the polynomial surface fitting degree d and length S of the series of threshold, the accuracy measures would be different.

In experiment, noise ratio β is set as 0.035 empirically, which means if the number of pixels of a filament candidate is less than 0.035 times the length (in number of pixels) of the radius of the solar disk, the filament candidates are treated as noise and removed from

the result. Morphological closing disk structure element radius P is set to be one percent of the length of the radius of the solar disk.



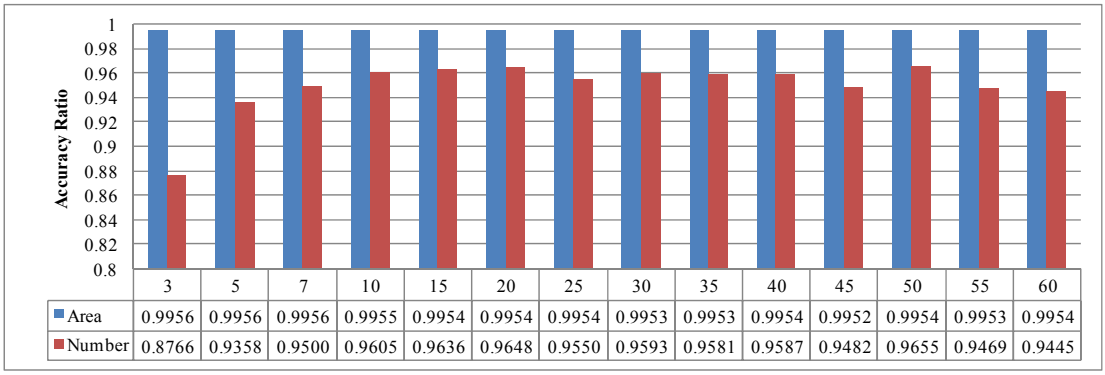
	0	1	2	3	4	5
Area	0.9950	0.9952	0.9953	0.9954	0.9952	0.9952
Number	0.8945	0.9358	0.9556	0.9648	0.9618	0.9630

Figure 2.17 Accuracy ratios with respect to different degrees of polynomial surface fitting for unbalanced luminance correction.

To measure the effect of the proposed unbalanced luminance correction method on the effect of solar filament detection, the degree d of polynomial surface fitting is varied and S is kept as 20. Figure 2.17 shows that the area-ratio and number-ratio changes related to d, where d equals zero indicating that no luminance correction is used. Results show that both area-ratio and number-ratio increase as d increases from zero to three, and then decrease. It is concluded that the third degree polynomial surface fitting works best for unbalanced luminance correction.

To measure the effect of the length of the series of threshold S on the effect of solar filament detection, the length S is varied. Figure 2.18 shows the area-ratio and number-ratio changes with respect to S. Results show that area-ratio does not change much, but number-ratio increases as S increases from 3 to 20. The number-ratio reaches

peak at 0.9648 when $S = 20$. When S increases to over 20, number-ratio fluctuates, but never over 0.96. This concludes that when $S = 20$, the proposed method performs the best.



	3	5	7	10	15	20	25	30	35	40	45	50	55	60
Area	0.9956	0.9956	0.9956	0.9955	0.9954	0.9954	0.9954	0.9953	0.9953	0.9954	0.9952	0.9954	0.9953	0.9954
Number	0.8766	0.9358	0.9500	0.9605	0.9636	0.9648	0.9550	0.9593	0.9581	0.9587	0.9482	0.9655	0.9469	0.9445

Figure 2.18 The accuracy ratio as a function of the length of the series of threshold.

Table 2.3 shows the performance (measured with filament area ratio and number ratio) of the proposed solar filament detection algorithm on Hα solar images obtained by four different solar observatories by using previously selected parameters $S = 20$ and $d = 3$. The algorithm performs well across the four observatories when measure by area ratio. It performs worst on solar images obtained by YNAO when measured by number ratio that is much lower than that of other three stations. It indicates that there are many tiny filaments are not detected from the solar images obtained by YNAO, which results from the bad quality of some of the solar images obtained by YNAO, such as one illustrated in Figure 2.19.

Table 2.3 Filament Detection Accuracy on Solar Images from Different Observatories

Observatory	BBSO	KANZ	OACT	YNAO
Area Ratio	0.9950	0.9964	0.9930	0.9991
Number Ratio	0.9514	0.9919	0.9717	0.7500

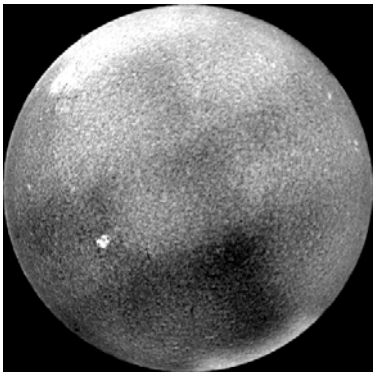


Figure 2.19 An Hα solar image obtained by YNAO on September 23, 2009 with bad quality.

Also conducted is the comparisons of the proposed solar filament detection algorithm on the same set of BBSO data which were used in [12]. The method in [12] was shown to outperform the previous two methods in [9, 10]. In experiment, 40 Big Bear Solar Observatory Hα images were randomly chosen from 2000 to 2010. The results show that the accuracy measured by filament area is 0.9445, and measured by filament number is 0.8240. Meanwhile, using our proposed method, the accuracy measured by filament area is 0.9949, and measured by filament number is 0.9342.

2.5.4 Summary

Experimental results show that: (1) The accuracy of the proposed automatic filament segmentation method is about 99% and 96% measured by area and by number of solar filament, respectively. (2) The best solar limb detection method is the combination of Median smooth filter and Roberts edge operator. (3) The third degree polynomial surface fitting is produces the best result for unbalanced luminance correction.

2.6 Application on Filament Tracking

Solar eruptions (flares, CMEs) are generally accompanied by solar filament (such as disappearances) [1]. Filament tracking is vitally important to understand solar activities. In this section, an automated filament tracking method is proposed built on the filament detection method proposed above.

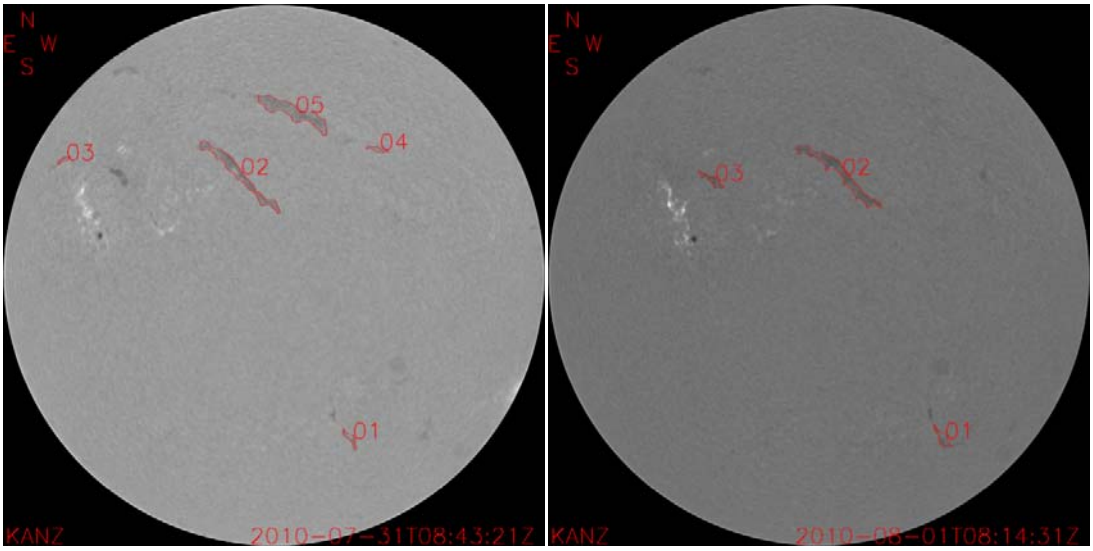


Figure 2.20 Hα solar images of two consecutive dates (Courtesy of KANZ).

Figure 2.20 illustrates Hα solar images of two consecutive dates (Jul. 31st, 2010 and Aug. 1st, 2010). Solar filaments presented in these two images (enclosed in red contours) are identified using the method presented in the preceding sections.

The challenge of solar filament tracking lies in two parts: First, a solar filament is not a rigid body. The shape of a solar filament is changing constantly. It may shrink, expand, and break into several pieces. Second, the movement of a solar filament is not regular. In general, a solar filament always moves from east to west due to solar rotation; however, it move differentially on the vertical direction [1].

A fuzzy association method for solar filament tracking is described in this section. Suppose that each filament can be accurately segmented using the method in the preceding sections. Then for each pixel of a filament, its rough location after a specific time can be figured out based on solar rotation. By comparing the location of a filament figured out by solar rotation and the location derived from observations, the association between two filaments can be established.

Suppose that there are two solar images $f_1(x,y)$ and $f_2(x,y)$ obtained by solar observatories at time t_1 and t_2. After the filament segmentation method proposed in the preceding sections is applied, there are m filaments segmented from $f_1(x,y)$ and n filaments segmented from $f_2(x,y)$. The m filaments segmented from $f_1(x,y)$ form a set $S_1 = \{\lambda_1, \lambda_2, ..., \lambda_m\}$, and the n filaments segmented from $f_2(x,y)$ form a set $S_2 = \{\gamma_1, \gamma_2, ..., \gamma_n\}$, where λ_i represents the set of pixels belonging to the i th filament in $f_1(x,y)$, and γ_i represents the set of pixels belonging to the i th filament in $f_2(x,y)$. For the purpose of filament tracking, each pixel of a filament is represented by its latitude x, longitude y and label l (the numbering of filament to which the pixels belongs). And thus λ_i ($i \in \{1,2,...,m\}$) itself is a set of triples. Assuming that the i th filament in image $f_1(x,y)$ is composed of p pixels, then $\lambda_i = \{< x_{i1}, y_{i1}, l_{i1} >, < x_{i2}, y_{i2}, l_{i2} >, ..., < x_{ip}, y_{ip}, l_{ip} >\}$. Similarly, assuming that the i th filament in image $f_2(x,y)$ is composed of q pixels, then $\gamma_i = \{< x_{i1}, y_{i1}, l_{i1} >, < x_{i2}, y_{i2}, l_{i2} >, ..., < x_{iq}, y_{iq}, l_{iq} >\}$.

According to differential rotation of the Sun[68], for a pixel at latitude x, its angular velocity (change of longitude y) in degrees per day is described by the following equation [69]:

$$\begin{aligned} &\omega = A + B\sin^2(x) + C\sin^4(x) \\ &A = 14.713 \quad (\pm 0.0491) \\ &B = -2.396 \quad (\pm 0.188) \\ &C = -1.787 \quad (\pm 0.253) \end{aligned} \tag{2.10}$$

Utilizing the equation above, the new longitude of each pixel can be figured out. Thus at time t_2, the m filaments segmented from $f_1(x,y)$ can be represented by $S_1' = \{\lambda_1', \lambda_2', ..., \lambda_m'\}$, where

$$\begin{aligned} \lambda_i' &= \{< x_{i1}, y_{i1}', l_{i1} >, < x_{i2}, y_{i2}', l_{i2} >, ..., < x_{ip}, y_{ip}', l_{ip} >\} \\ &= \{< x_{i1}, y_{i1} + (t_2 - t_1)\omega(x_{i1}), l_{i1} >, < x_{i2}, y_{i2} + (t_2 - t_1)\omega(x_{i2}), l_{i2} >, ..., \\ &< x_{ip}, y_{ip} + (t_2 - t_1)\omega(x_{ip}), l_{ip} >\} \end{aligned} \tag{2.11}$$

where $\omega(x_{ij})$ is the angular velocity of the j th pixels of the i th filament segmented from $f_1(x,y)$.

Given $S_1' = \{\lambda_1', \lambda_2', ..., \lambda_m'\}$ and $S_2 = \{\gamma_1, \gamma_2, ..., \gamma_n\}$, the association between the filaments in solar images $f_1(x,y)$ and $f_2(x,y)$ can be figured out. Assume i th filament of $f_1(x,y)$ is composed of p pixels and j th filament of $f_2(x,y)$ is composed of q pixels,

Figure 2.21 illustrates a procedure which can be used to find out whether the i th filament of $f_1(x,y)$ is associated with the j th filament of $f_2(x,y)$.

```
initialize counter δ=0
for u from 1 to p
    for v from 1 to q
        if the distance between (x_iu, y_iu') and (x_jv, y_jv) is less than a
threshold ζ
            increment counter δ by 1
        endif
    endfor
endfor
```

Figure 2.21 A procedure for filament association.

If the counter δ is greater than 0 after the running of procedure above, then i th filament of $f_1(x,y)$ is associated with the j th filament of $f_2(x,y)$. The value of counter denotes the rate of overlapping between the two filaments. Running the procedure above on each pair of the filaments (one from $f_1(x,y)$ and another from $f_2(x,y)$), the association between the filaments of $f_1(x,y)$ and $f_2(x,y)$ can be figured out.

Let a time series of n solar images be $f_1(x,y), f_2(x,y),\ldots, f_n(x,y)$ obtained at time t_1, t_2,…, t_n. For each pair of images $f_i(x,y)$ and $f_{i+1}(x,y)$ ($i \in \{1,2,...,n-1\}$), the

method above can be used to find out the association of filaments among them, and thus it is trivial to track the filaments during the time from t_1 and t_n.

Figure 2.22 illustrates the result of filament tracking from Aug 20 to Sep 1 2003. The x-axis shows the date information; the y-axis shows the area information (in number of pixels). The lines in the figure illustrate the change of areas of filaments with respect to time. Different filaments are illustrated using different colors. The numbering on the line illustrates the label of the filament on the solar disk at the designated date. As it can be seen, filament number 2 on August 28 is split into two filaments (number 2 and number 4) on August 29. Filament number 4 and 5 on August 26 merged to filament number 7 on August 27. It is also noted that the missing of solar images on August 25 and August 31 does not compromise the filament tracking.

Figure 2.23 illustrates the result of filament segmentation from August 20 to September 1, 2003. The boundaries of the segmented filament are marked in red color. Each filament is also marked with a label (numbering), which corresponds to the label in Figure 2.22. By comparing the solar images in Figure 2.23 and the lines in Figure 2.22, the filament tracking method is very successful.

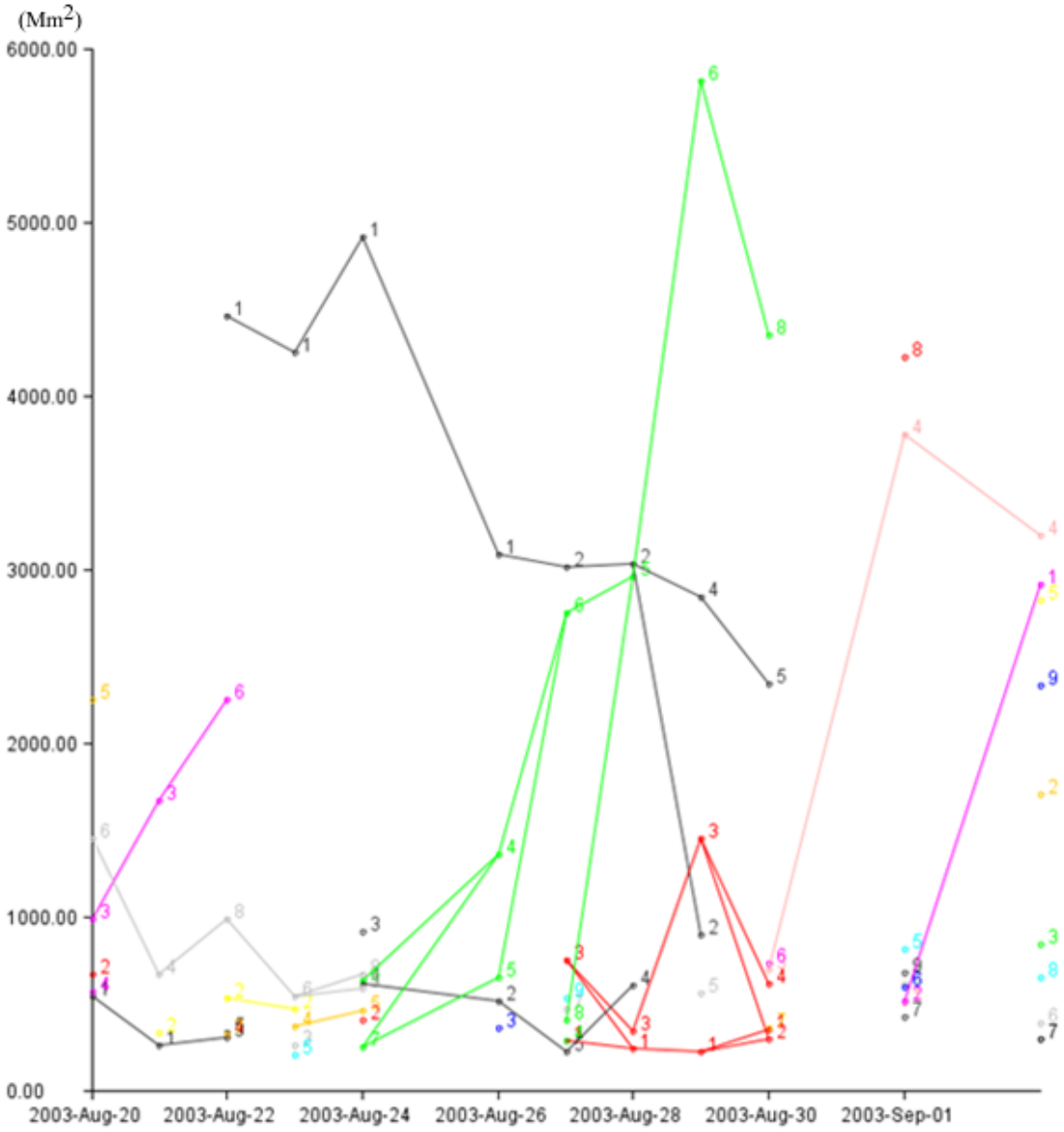


Figure 2.22 Illustration of filament tracking from Aug. 20 to Sep. 3, 2003. The figure illustrates the change of areas of filaments with respect to time. Each dot in the figure represents a filament. The x-coordinate of each dot shows the date of the filament, while the y-coordinate of each dot shows the area of the filament. Dots of different dates are connected with lines if they belong to the same filament appearing on different dates. The number near each dot is the labeling of the corresponding filament, which can be found out in Figure 2.23.

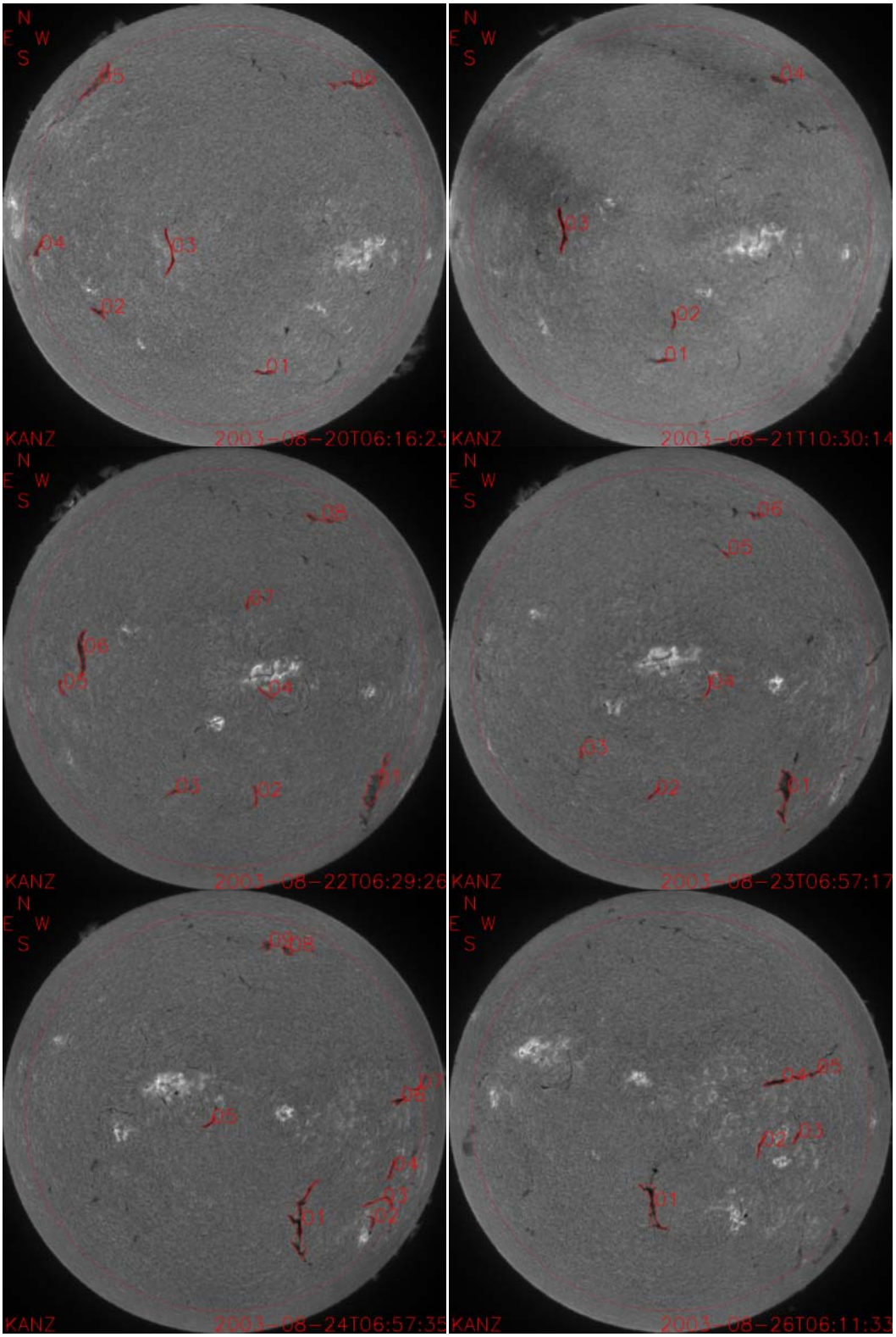


Figure 2.23 Illustration the results of filament segmentation from Aug. 20 to Sep. 3, 2003. 12 solar images are illustrated. The boundaries of solar filaments presented in these images are over-plotted. A number is assigned to each filament.

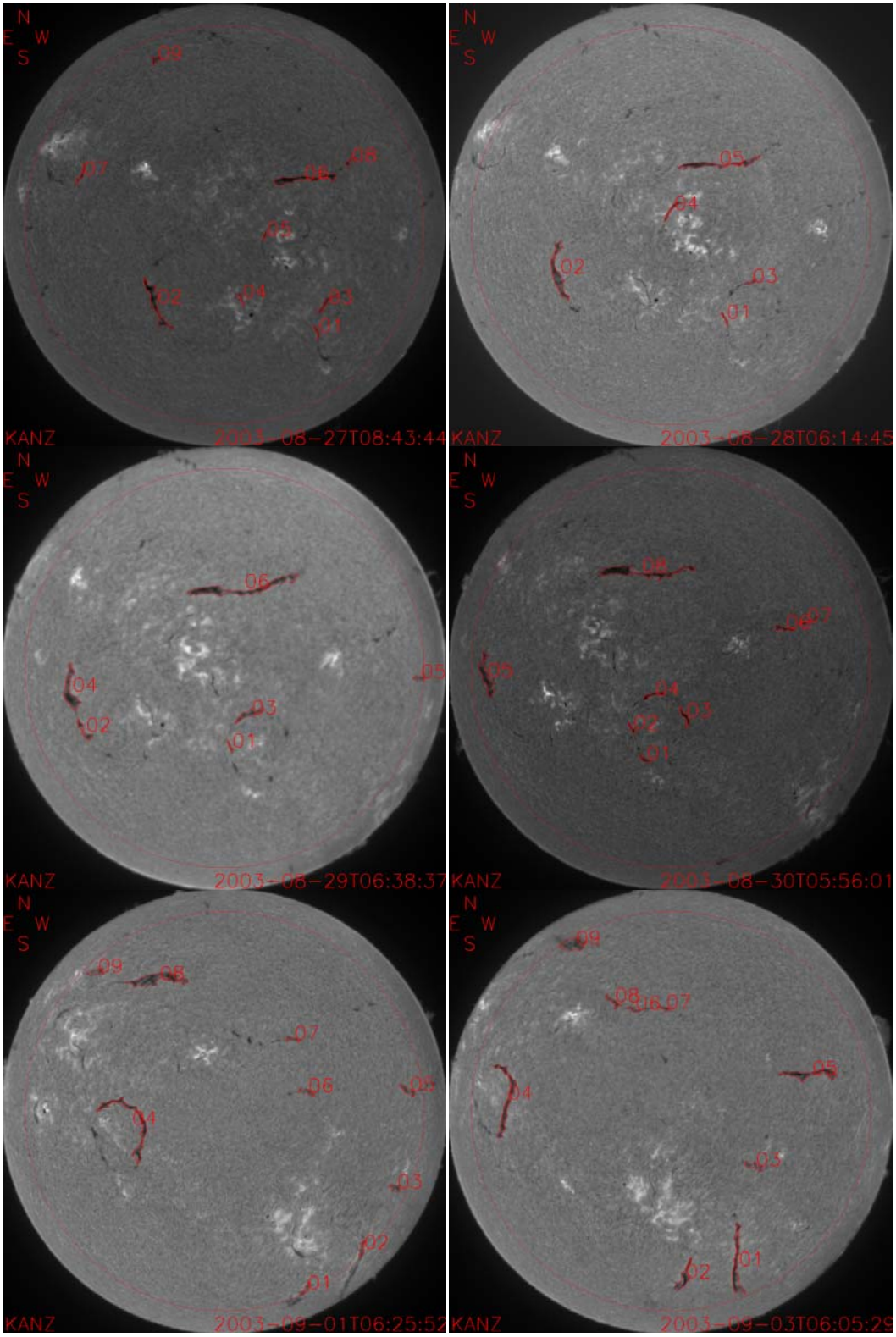


Figure 2.23 (Continued)

2.7 Summary

In this chapter, a solar filament segmentation and characterization algorithm is proposed, which aims to automatically detect and characterize solar filaments in Hα solar images obtained from different solar observatories. Experimental results on 125 solar images captured by four different solar observatories show that the accuracy of the proposed method is more than 99% and 96% (as illustrated in Figure 2.18 when S=20) as measured by area and by number of solar filaments, respectively.

For filament characterization (such as heliographic centroid location), the center location and radius of the solar disks are identified by using a two-stage Hough circle detection algorithm to mitigate the limitation imposed by traditional Hough circle detector. Experimental results show that the quality measure of the edge points obtained by median filter with Roberts edge operator can reach 74%.

The accuracy ratio changes as the length of the series of the threshold values S changes. On the one hand, increment of S can make the selected threshold value close to the optimal threshold value; on the other hand, when S becomes too large, the result is vulnerable to noise. The idea can be explained by Figure 2.24. When the length S is moderate, the average high pass filter response curve is smooth (shown as the solid curve), whose peak is the optimal threshold value. When the length S is too small, the curve (shown as dotted curve) is too smooth to get an accurate threshold value. Conversely, when the length S is too large, the curve (shown as dashed curve) is too rough, and its peak may deviate from the real optimal value.

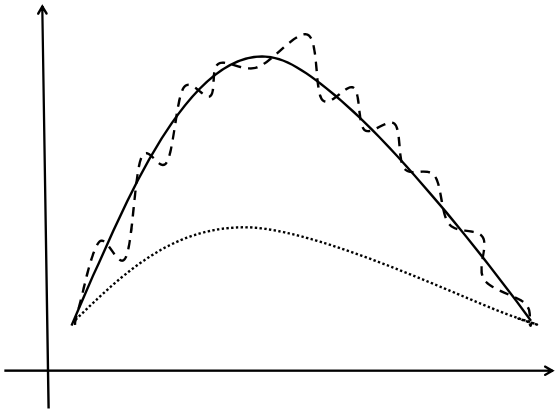


Figure 2.24 Illustration of average high pass filter response curve within different regions.

An application of the proposed filament segmentation method to filament tracking is illustrated and preliminary results show that the performance is very good. The area change of filaments is recorded, and splitting and merging of filaments are also recorded, which presents a whole picture of the life span of filaments during a time period.

CHAPTER 3

AUTOMATED TRACING OF CHROMOSPHERIC FIBRIL

3.1 Introduction

Our understanding of the energy storage and release mechanism of solar eruptive events is strongly dependent on knowledge of their magnetic field configurations. Unfortunately, although recent advances in near-infrared spectropolarimetric instrumentation [70-72] and broadband imaging spectroscopy with solar radio telescope [73, 74] show great promise, current technologies have yet to succeed to a level that makes reliable and routine measurements of chromospheric and coronal magnetic fields.

On the other hand, chromospheric fibrils seen in the Hα central line are threads of mass that appear in abundance throughout the field-of-view (FOV) of Hα filtergrams. It is generally believed that fibrils are magnetic field-aligned, primarily due to the reason that the high electrical conductivity of the solar atmosphere “freezes" the ionized mass in magnetic field lines and prevents them from diffusing across the lines. Very recently, [13] test this common notion for the first time by comparing the orientation of fibrils to the azimuth of chromospheric magnetic fields obtained by spectropolarimetric measurements of Ca II lines, and found a general alignment as well as some discrepancy between the two directions. [13] ascribe the discrepancy to either the difference in formation height or the time lag between the fibril and magnetic field measurements.

Since it is mostly true that fibrils are oriented along the magnetic field direction theoretically and observationally, it would be reasonable to adopt chromospheric fibrils as a surface tracer of chromospheric magnetic fields. A method that automatically segments fibrils from Hα images and further identifies their orientation is presented. This method is

applied to Hα images of active region NOAA 9661 on October 19, 2001 and active region NOAA 11092 on August 3, 2010.

3.2 Segmentation and Modeling of Chromospheric Fibril

Chromospheric fibrils seen in Hα images are segmented by a threshold-based method, and then modeled with polynomial-curves. The fibril tracing steps are as follows:

First, an Hα image $f(x,y)$ is smoothed by convolution with a 2-dimensional Gaussian filter $G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}$, where x and y specifies the size of the Gaussian filter and σ is the standard deviation. The difference image between the original image and the smoothed image $f''(x,y)=f(x,y)-f'(x,y)$ is figured out. Clearly, the difference image $f''(x,y)$ demonstrates contrast enhancements and thus is better for the fibril segmentation in comparison with the original image $f(x,y)$.

Second, the fibrils are segmented from $f''(x,y)$ with the threshold $t=\mu\ \text{median}\left(f''\left(x,y\right)\right)$, where median function computes the median value of $f''(x,y)$ and $\mu=\frac{7}{8}$. Since fibrils are dark features over the relatively bright solar disk in Hα images, those pixels with brightness below the threshold t are considered as candidate elements constituting fibrils. The result after this step is a binary image $k\left(x,y\right)$ where each pixel is either 1 or 0.

Third, union-find algorithm [75] is used to group adjacent pixels (8-neighbor) in $k\left(x,y\right)$ [76] to form fibril candidates. Since small fibril candidates (e.g., the total number of pixels is less than 100) tend to appear like dots and hence fail to show orientation, these

small fibril candidates are removed from $k(x,y)$. In addition, sunspots that are presented in the segmentation result $k(x,y)$ need to be removed. Different from fibrils, sunspots are round structures. A shape descriptor $q=\frac{p}{4\pi a}$ is designed, where p and a is the perimeter and area of a fibril candidate respectively. Because the geometric figure of maximum area and given perimeter is a circle. The shape descriptor q of a disk structure is one. And shape descriptor of a structure decreases as the shape of a structure change from a round structure to a elongated string-like structure. And thus it is possible to remove sunspots by removing those fibril candidates whose shape descriptors are greater than a threshold value. It is found out 0.7 is a good threshold value to differentiate sunspots from fibrils. The result, $k'(x,y)$, after removing small fibril candidates and sunspots, contains fibrils only. The segmented, thread-like fibrils sketch out the basic configuration of the chromospheric magnetic field and will be modeled in the next step.

Next, each fibril is modeled by a $m-1$ degree polynomial curve $h(x,\boldsymbol{\beta})$. Suppose that a fibril is consists of n data points $(x_i,y_i), i=1,2,\cdots,n$, where x_i and y_i is the horizontal and vertical coordinates of the ith data points. Least squares fitting is used to find out the m parameters held in the vector $\boldsymbol{\beta}$. The least squares method finds the optimum value of $\boldsymbol{\beta}$ when the sum, S, of squared residuals is a minimum. A residual, r, is defined as the difference between the value predicted by the model and the actual value of the vertical coordinates. The minimization problem can be expressed as following [77]:

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} S = \min\limits_{\boldsymbol{\beta}} \sum\limits_{i=1}^{n} r_i^2 \\ r_i = y_i - h\left(x_i, \boldsymbol{\beta}\right) \end{cases} \tag{3.1}$$

The minimum of the sum of squares is found by setting the gradient to zeros as following:

$$\frac{\partial S}{\partial \beta_j} = \sum_i \frac{r_i^2}{\partial \beta_j} = 2\sum_i r_i \frac{\partial r_i}{\partial \beta_j} = -2\sum_i r_i \frac{\partial h(x_i, \boldsymbol{\beta})}{\partial \beta_j} = 0,\ j = 1, 2, \cdots, m \tag{3.2}$$

Because the polynomial curve $h(x, \boldsymbol{\beta})$ comprises a linear combination of the parameters $\boldsymbol{\beta}$, i.e.

$$h(x, \boldsymbol{\beta}) = \sum_{j=1}^{m} \phi_j\left(x\right)\beta_j \tag{3.3}$$

Letting

$$X_{ij} = \frac{\partial h\left(x_i, \boldsymbol{\beta}\right)}{\partial \beta_j} = \phi_j\left(x_i\right), i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m \tag{3.4}$$

It can be derived that

$$\frac{\partial S}{\partial \beta_j} = -2\sum_{i=1}^{n} r_i \frac{\partial h(x_i, \boldsymbol{\beta})}{\partial \beta_j} = -2\sum_{i=1}^{n} \left(y_i - h(x_i, \boldsymbol{\beta})\right) X_{ij} = -2\sum_{i=1}^{n} X_{ij}\left(y_i - \sum_{k=1}^{m} X_{ik}\beta_k\right) \tag{3.5}$$

Thus if $\hat{\boldsymbol{\beta}}$ minimizes S, then

$$-2\sum_{i=1}^{n} X_{ij}\left(y_i - \sum_{k=1}^{m} X_{ik}\hat{\beta}_k\right) = 0 \tag{3.6}$$

Rearrangement would gives that

$$\sum_{i=1}^{n}\sum_{k=1}^{m} X_{ij}X_{ik}\hat{\beta}_k = \sum_{i=1}^{n} X_{ij}y_i, j = 1,2,\cdots,m \tag{3.7}$$

Written in matrix notation, the equation above would be

$$\left(\mathbf{X^T X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X^T y} \tag{3.8}$$

where $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$. The solution of the equation above yields the optimal parameter values $\hat{\boldsymbol{\beta}} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T y}$.

Finally, at each particular point of the polynomial curve, evaluating the derivative of the curve yields the slope of the tangent and hence the orientation angle with respect to the x-axis.

In the description above, each fibril is modeled as a polynomial function in x to simplify the description of the algorithm. In reality, a fibril is modeled as a polynomial function in y if it is nearly vertical.

3.3 Experimental Results

The proposed fibrils segmentation and modeling method are applied to two Hα images. These two Hα images are obtained by BBSO on October 19, 2001 and Osservatorio Astrofisico di Arcetri on August 3, 2010 respectively.

Figure 3.1 illustrates the segmentation and modeling of fibrils on Hα image obtained by BBSO. Panel 1 of Figure 3.1 illustrates the original BBSO Hα image taken at 16:04 UT on October 19, 2001. Panel 2 illustrates the difference image between the enhanced image which is the difference image between original and smoothed image. Panel 3 illustrates fibrils candidates after performing image thresholding. Panel 4 illustrates fibrils after removing small fibrils and sunspots. Panel 5 illustrates the second-degree-polynomial modeling of fibrils (red curves), overlaid on the original Hα image. Panel 6 illustrates the orientation of the modeled fibrils. The value of orientation angle is indicated by the color scale bar.

Figure 3.2 illustrates the segmentation and modeling of fibrils on Hα image obtained by Osservatorio Astrofisico di Arcetri. Panel 1 of Figure 3.2 illustrates the original Hα image taken on August 3, 2010. Panel 2 illustrates the difference image between the enhanced image which is the difference image between original and smoothed image. Panel 3 illustrates fibrils candidates after performing image thresholding. Panel 4 illustrates fibrils after removing small fibrils and sunspots. Panel 5 illustrates the third-degree-polynomial modeling of fibrils (red curves), overlaid on the original Hα image. Panel 6 illustrates the orientation of the modeled fibrils. The value of orientation angle is indicated by the color scale bar.

It can be seen that most fibrils are segmented correctly. All major fibrils are presented in the segmentation result, although some thinner fibrils are undetected. Second-degree-polynomial curves are used for modeling the fibrils in Figure 3.1 but third-degree-polynomial curves are used in Figure 3.2. Because fibrils in Figure 3.1 is shorter and smoother, third-degree-polynomial modeling would cause oscillations; while first-degree-polynomial modeling, which produces straight lines, would not capture the shapes of most fibrils. For fibrils in Figure 3.2, third-degree-polynomial modeling produces the best results.

3.4 Summary

A fibril segmentation and modeling method is presented in this chapter. Since it is mostly true that fibrils are oriented along the magnetic field direction theoretically and observationally, it would be reasonable to adopt fibrils as a surface tracer of chromospheric magnetic fields, which helps in our understanding of the energy storage and release mechanism of solar eruptive events.

Image processing techniques such as image enhancement, image segmentation, and union-find are used to segment fibrils from Hα images. Least squares curve fitting is used to model segmented fibrils. Experimental results show that the proposed method is very successful in segmentation and modeling of most fibrils, especially major fibrils.

For future research, the least square fitting of fibrils can be improved by introducing optimization mechanism to search for a good balance between smoothness (low order polynomial fitting) and accuracy (high order polynomial fitting, but can cause oscillations).

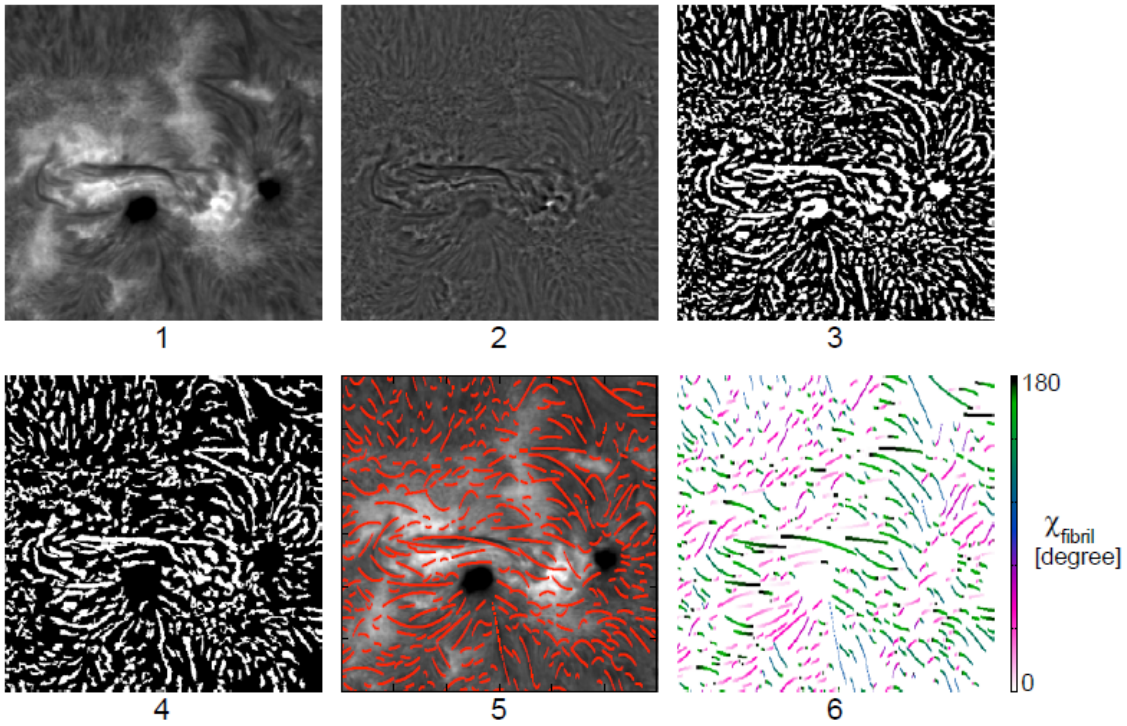


Figure 3.1 Segmentation and modeling of Hα fibrils. Panel 1: the original BBSO Hα image obtained on October 19, 2001; Panel 2: the difference image between the original and a smoothed image; Panel 3: the segmented fibril candidates after image thresholding; Panel 4: the same as Panel 3, except that the segmented pieces shown in Panel 3 are grouped with the union-find algorithm and small groups and sunspots are removed from the image; Panel 5: the second-degree-polynomial modeling of fibrils (red curves), overlaid on the original Hα image; Panel 6: the orientation of fibrils. The value of orientation angle is indicated by the color scale bar. The field-of-view (FOV) is 240×240″, corresponding to 174×174 Mm.

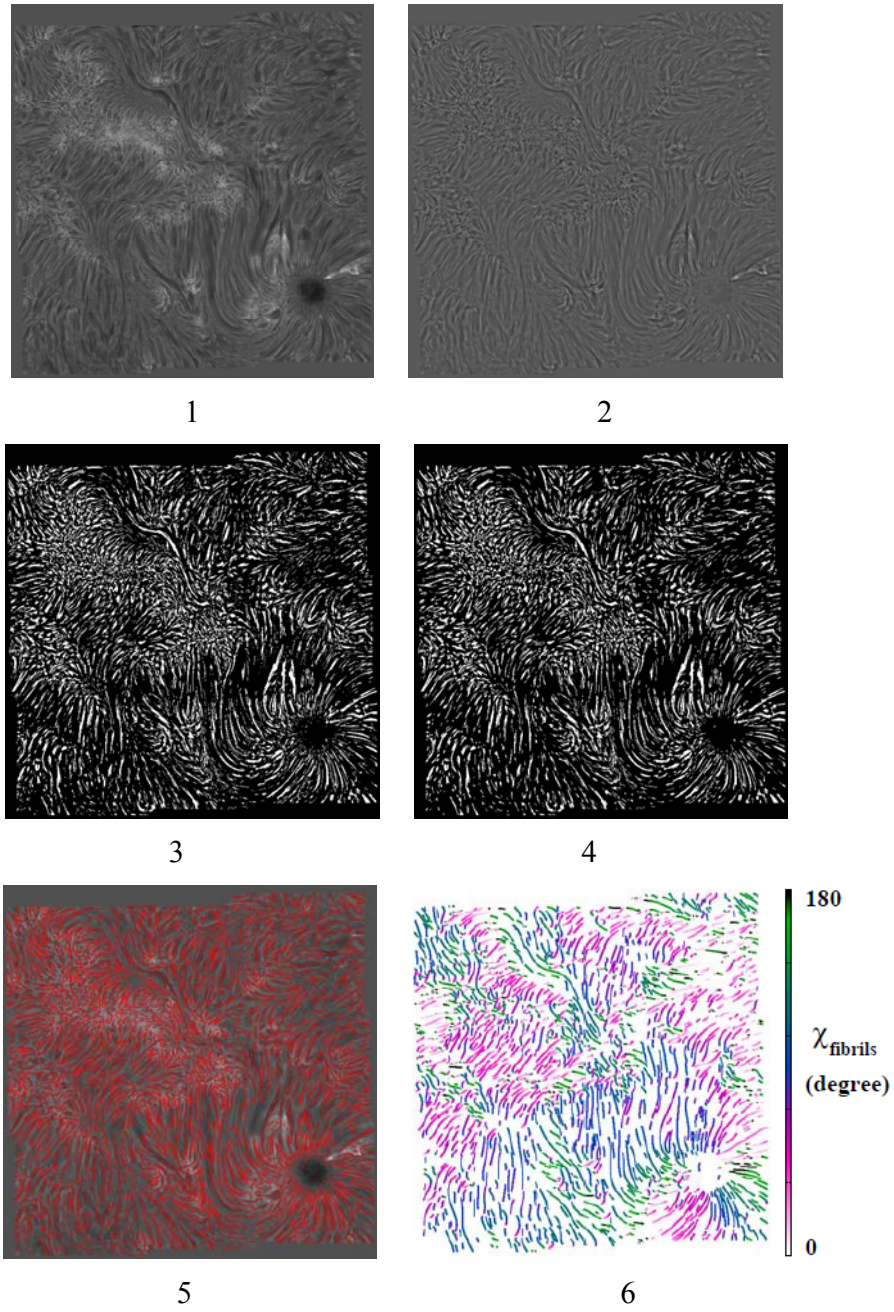


Figure 3.2 Segmentation and modeling of Hα fibrils. Panel 1: the original Hα image obtained by Osservatorio Astrofisico di Arcetri on August 3, 2010; Panel 2: the difference image between the original and a smoothed image; Panel 3: the segmented fibril candidates after image thresholding; Panel 4: the same as Panel 3, except that the segmented pieces shown in Panel 3 are grouped with the union-find algorithm and small groups and sunspots are removed from the image; Panel 5: the third-degree-polynomial modeling of fibrils (red curves), overlaid on the original Hα image; Panel 6: the orientation of fibrils. The value of orientation angle is indicated by the color scale bar.

CHAPTER 4

AUTOMATED FLARE FORECASTING USING A STATISTICAL LEARNING TECHNIQUE

4.1 Introduction

The sudden and intense release of energy stored in solar magnetic fields generates solar flares [78], which can have a significant impact on the near earth space environment (so called space weather). The development of fully automatic programs to detect [79, 80] and forecast flares is regarded as one of the most important tasks to process the large amount of data accurately and efficiently.

At present, a number of different flare forecasting approaches and systems have been developed based on photospheric magnetic field observations or sunspot-group characteristics. For instance, "Theophrastus," a system developed by the Space Weather Prediction Center of NOAA, is mainly based on the correlation between solar flare production and sunspot-group classification [81]. At Big Bear Solar Observatory, Gallagher *et al.* [82] used the historical average of flare numbers according to the McIntosh classification to develop a solar flare prediction system which estimated the probabilities for each active region to produce C-, M-, or X-class flares. Barnes and Leka [83] adopted discriminant analysis to accomplish solar flare forecasting within 24 hours using a large combination of vector magnetic field measurements obtained by the University of Hawaii Imaging Vector Magnetograph. Li *et al.* [84] proposed a solar flare forecasting method based on support vector machines in which the sunspot area, the sunspot magnetic class, the McIntosh class of the sunspot group and the 10 cm solar radio flux were chosen as precursors. Georgoulis and Rust [85] defined a new measurement called the effective

connected magnetic field, and their experimental results, based on 298 active regions during a 10 year period of solar cycle 23, showed that this measure was an efficient flare-forecasting criterion. Qahwaji and Colak [86] put forward a short-term solar flare prediction method using machine learning and sunspot associations, in which the authors had compared the performance of the proposed method with two other machine learning algorithms.

Different from the approaches mentioned above, Wheatland [87] designed a Bayesian approach to solar flare prediction in which only the statistics of flare events was used as predictors; however, this approach has not been tested on a large data set.

In this chapter, a new method is put forward for the automatic forecasting of the occurrence of solar flares over 24 hours following the time when a magnetogram is presented. The method is a continuation and extension of the method proposed by Song *et al*. [24], which has some limitations in forecasting X-class flares. The proposed method is split into two cascading steps. In the first step, logistic regression is used to map three magnetic parameters of each active region into four probabilities; support vector machine classifier is then utilized to map the four probabilities onto a binary label which is the final output. Experimental results illustrate that the proposed method performs better for X-class flare forecasting.

The chapter is organized as follows. The definitions of the predictive variables (i.e., three magnetic parameters) used in this study are introduced in Section 4.2. The proposed flare forecasting method is described in Section 4.3. Experimental results are shown in Section 4.4. Finally, a conclusion is drawn in Section 4.5.

4.2 Data Description

4.2.1 Predictive Variables

To be consistent with the work of Song *et al*. [24], the same predictive variables are used. The predictive variables of Song *et al*. [24] are composed of:

1. Total unsigned magnetic flux, T_{flux}, which is the integration of pixel intensity over the area of an active region,

$$T_{flux} = \iint |B_Z| dxdy \tag{4.1}$$

where Bz is the pixel intensity of MDI magnetograms.

2. Length of the strong-gradient magnetic polarity inversion line, L_{gpi}, which was first studied by Falconer *et al*. [88] as a measure to predict coronal mass ejections. Jing *et al*. [20] illustrated the correlation between L_{gpi} and flare productivity of active regions. As illustrated in Song *et al*. [24], L_{gpi} is the total number of pixels on which the gradient $|\nabla \perp B_z|$ is greater than a threshold, which is 50G Mm^{-1} as chosen by Song *et al*. [24]. The definition of $|\nabla \perp B_z|$ is as follows:

$$|\nabla \perp B_z| = \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right]^{1/2} \tag{4.2}$$

3. Total magnetic energy dissipation, E_{diss}, proposed by Abramenko *et al*. [89], was also studied by (Jing *et al*.[20]; Song *et al*.[24]) in exploring its correlation between flare productivity of active regions. According to Abramenko *et al*.[89],

$$\mathrm{E}_{\mathrm{diss}} = \iint 4\left[\left(\frac{dB_z}{dx}\right)^2 + \left(\frac{dB_z}{dy}\right)^2\right] + 2\left(\frac{dB_z}{dx} + \frac{dB_z}{dy}\right)^2 dxdy \tag{4.3}$$

where the integration is performed over the area of an active region.

These parameters are chosen because: (1) all three can be derived from the line-of-sight magnetograms; and (2) all three moderately correlate with the flare productivity of active regions and show their forecasting utility in the previous study by Jing *et al*. [20] and Song *et al*. [24].

4.2.2 Data Collection

The three magnetic parameters introduced above were derived from the magnetograms produced by the Michelson Doppler Imager (MDI), which is an instrument onboard the Solar and Heliospheric Observatory (SOHO).

This study uses the same dataset as what was used by Song *et al*. [24], which focuses on active regions between 1996 and 2005. It covers almost the entire solar cycle 23 which peaked in 2001. A total of 230 sample active regions were selected using the following criteria: (1) the center location of an active region is close to the solar disk center (within ± 40 degrees in longitude and ± 40 degrees in latitude); (2) the MDI full disk magnetograms are available; (3) since an active region may appear on the solar surface for

a few days, it is treated as a different sample on different dates; (4) the first magnetogram of the 15 magnetograms taken by MDI each day is chosen.

4.2.3 Correlation between Magnetic Parameters and Flare Productivity

Using the same criteria as [24], active regions are categorized into four levels according to the most powerful flare they produced: an active region is classified as level-0 if it is flaring-quiet or only produces A and/or B class flares; an active region is classified as level-1 if it produces at least one C-class flare but no M- or X- class flares; Level-2 corresponds to those active regions which produce at least one M-class flare but no X-class flares; Level-3 corresponds to those active regions which produce at least one X class flare.

Figures 4.1, 4.2 and 4.3 illustrate the histograms of the length of the strong gradient inversion line, total unsigned flux and energy dissipation. Please note the values are scaled to 0 and 1, and the unit is shown below each graph. The height of a bar denotes the number of samples whose corresponding parameters are within some range. Within each range, different colored bars are used to differentiate the samples into different levels.

For example, the height of the blue bar in Figure 4.3 within range 0 and 0.1 is 39, meaning that there are 39 level 0 samples whose energy dissipation is within the range 0 and 3.78×108 erg cm^{-3}. As it can be seen, the blue bars (which correspond to level 0 samples) are mainly distributed in the lower ranges, and their heights decrease as the values increase. The red bars (which correspond to level 3 samples) can reach higher ranges, which coincide with our observations that samples with higher values of these parameters are more likely to produce X-class flares.

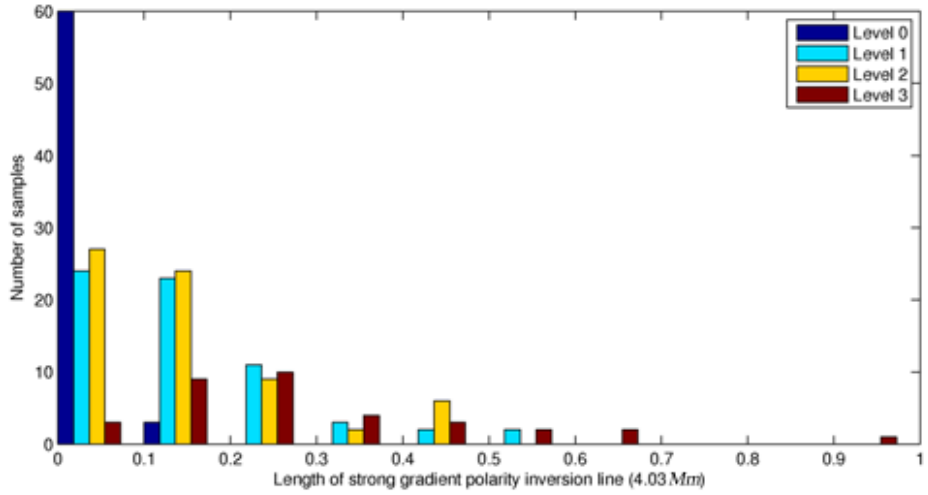


Figure 4.1 Histogram of the first parameter for the four different levels.

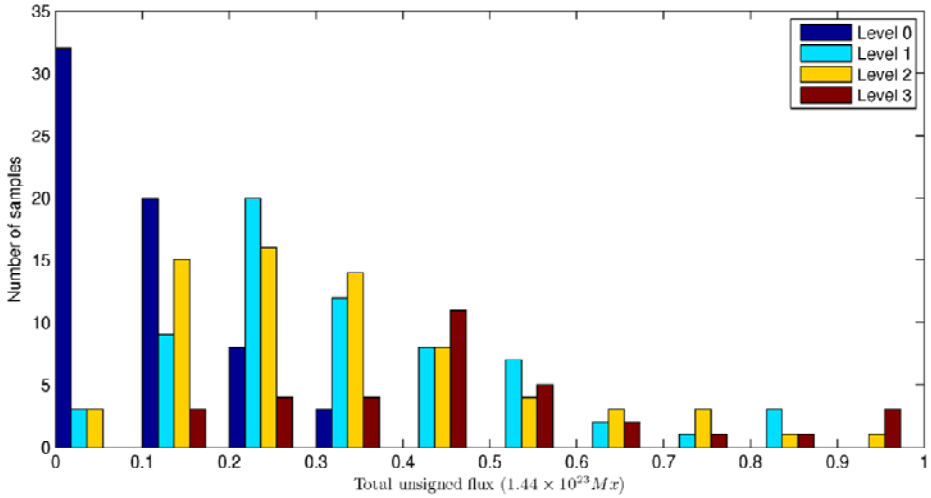


Figure 4.2 Histogram of the second parameter for the four different levels.

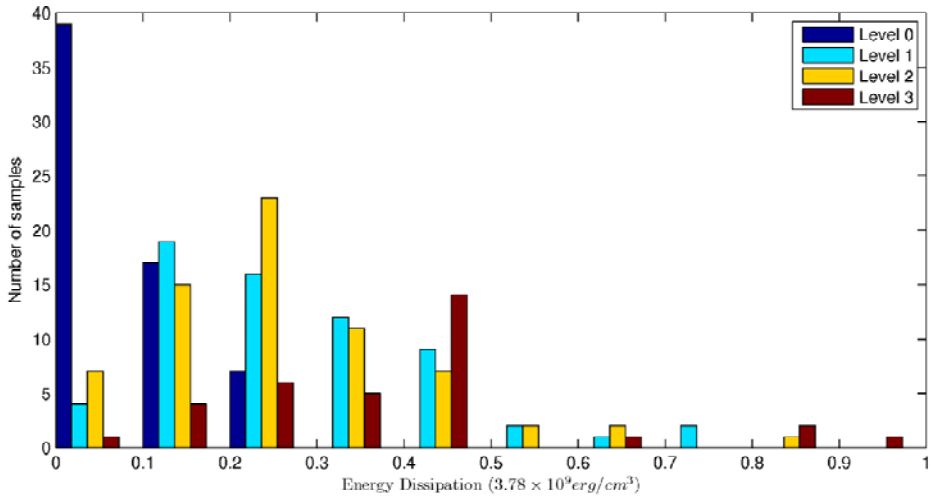


Figure 4.3 Histogram of the third parameter for the four different levels.

Table 4.1 Mean Value and Standard Deviation of Predictive Parameters

Active Region Level	Number of Active Regions	L_{gpi} (Mm)		T_{flux} ($10^{22}\,Mx$)		E_{diss} ($10^{8}\,erg\ cm^{-3}$)	
		Mean	Deviation	Mean	Deviation	Mean	Deviation
3	34	118.74	79.88	7.02	3.15	15.38	7.76
2	68	64.28	46.79	5.03	2.72	10.58	5.59
1	65	62.12	46.61	4.95	2.86	10.47	5.88
0	63	10.84	15.19	1.72	1.19	3.67	2.58

For the 230 active regions in the dataset, the correlations between magnetic parameters and flare productivity are summarized in Table 4.1. Table 4.1 shows that the mean value of the length of the strong-gradient magnetic polarity inversion line of the 34 level-3 active regions is 118.74, which is much larger than that of the 68 level-2 active regions (64.28). The mean value of the length of the strong-gradient magnetic polarity inversion line of 68 level-2 active regions is 64.28, which is slightly larger than that of the 65 level-1 active regions (62.12). The mean value of the length of the strong-gradient magnetic polarity inversion line of 63 level-0 active regions is 10.84, which is much less

than that of other levels of active regions. For total unsigned magnetic flux and total magnetic energy dissipation, the same kind of trend follows. However, since the deviation is large (almost half of the mean values), it is impossible to do precise flare forecasting based on those parameters.

Based on the correlations described above, statistical and machine learning methods are utilized to perform flare forecasting.

4.3 Forecasting Method

In previous studies, there are mainly two types of flare forecasting methods. The first type is based on pattern recognition, such as a Support Vector Machine-based (SVM-based) method [23]. During this kind of analysis, some predictive parameters of a given active region are extracted, and then the predictive parameters are fed into a trained classifier. The output of the classifier (usually a label indicating which class of flare is likely to occur) is the final forecasting result. The disadvantage of this type is that the output is only a label, which does not provide information on how much confidence can be placed on each forecast. For example (see Figure 4.4), both sample A and sample B will be classified as the same class, but obviously it is more confident to believe that B belongs to this class than A, because A is on the boundary. However, because the output of SVM is only a label, that kind of information is not presented.

The second type is based on probability analysis, such as ordinal logistic regression[24]. During this kind of analysis, some predictive parameters of a given active region are extracted, and then those predictive parameters are fed into a trained statistical model, and the output of the model is the probability that a flare event will occur. Of course, using a threshold value (generally 0.5), the probability can be converted into a binary

forecast. However, it is not an easy job to choose a good threshold value, and the de facto standard threshold (0.5) is not always the best, as illustrated in [24], where the authors chose 0.25 as the threshold for X-class flare prediction.

In this study, the proposed method is split into two steps (see Figure 4.5). In the first step, ordinal logistic regression is utilized to map the input (three predictive parameters of a given active region) to four outputs (the probabilities of the given active region belonging to each of the four levels). Secondly, the four outputs are fed into a support vector machine; the output of the support vector machine tells whether the given active region belongs to one level or not.

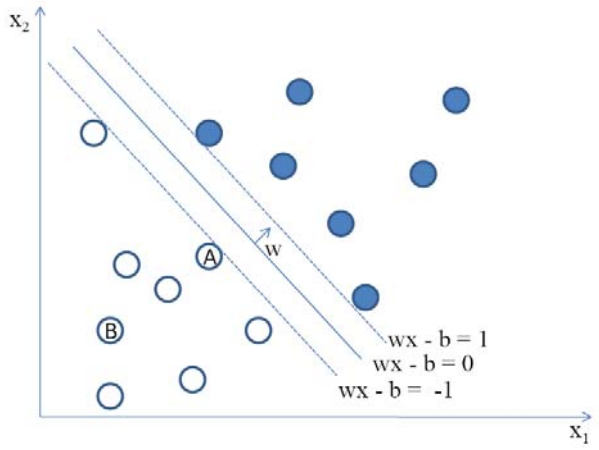


Figure 4.4 An illustration of the support vector machine classifier.

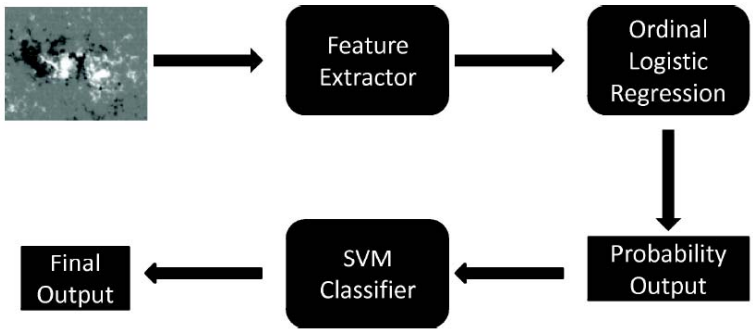


Figure 4.5 The workflow diagram of the proposed forecasting system.

Generally, the first step is enough for a flare forecasting system. The purpose of the second step is three fold. First, it is hard to assess the performance of the first step since the outputs are probabilities instead of a definite answer. Secondly, users sometimes want a definite answer instead of a probability. Thirdly, the outputs of the second step can be used to compare with other research whose outputs are only binary labels.

4.3.1 Probability Prediction Using Ordinal Logistic Regression

Used for Bernoulli-distributed dependent variables, logistic regression is a generalized linear model that uses the logit as its link function [90]. One common application of logistic regression is to estimate the probability of the occurrence of an event from predictive variables. Logistic regression is used to map predictive variables into probabilities of the occurrence of flares by [24, 91]. The comparison made by [24] shows that their forecasting results are better than those of the Solar Data Analysis Center and NOAA's Space Weather Prediction Center, which illustrates the usefulness of logistic regression in flaring probability estimation.

Suppose that the data in a dataset belong to L ordered levels and $P(D=g)$ is the probability that an event which belongs to level g would occur given predictive variables $\boldsymbol{X}$, then, according to [92],

$$\begin{aligned} &P\left(D=g\right)=P\left(D\geq g\right)-P\left(D\geq g+1\right), \\ &P\left(D\geq g\right)=\frac{1}{1+e^{-\left(\alpha_g+\beta^T\mathbf{x}\right)}}, \\ &g=1,2,3,...,L. \end{aligned} \tag{4.4}$$

Given a training dataset composed of predictive variables and response category pairs, the parameters $\alpha_g, g = 1,2,3,...,L$ and β in the above equation can be calculated using maximum likelihood estimation [93].

The application of ordinal logistic regression to flare forecasting is as follows:

1. Training: The training data contain several samples; each sample is composed of three photospheric magnetic features of an active region and the level of the given active region.
2. Forecasting: Using the ordinal logistic regression model, for a given active region, at first, figure out its three photospheric magnetic features, and then feed these three variables into the model. The output of the model contains four elements, which correspond to the probabilities that the given active region belongs to level 0, 1, 2, or 3.

4.3.2 Binary Forecasting Using Support Vector Machines

An SVM is a supervised learning method used for classification[47], whose principle is to minimize the structural risk [94]. An SVM tries to find a plane in an n-dimensional space that separates input data into two classes. The larger the distance from the plane to the two different classes of data points in the n-dimensional space, the smaller the classification error [48].

Given training vectors $X_i \in R^d, i = 1,2,...,n$ in two classes labeled by a vector $y \in R^n$ where $y_i = \{-1,1\}, i = 1,2,...,n$. The training of a support vector machine is equivalent to solving the following optimization problem [95]:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}}\left(\frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha}-\mathbf{e}^T\boldsymbol{\alpha}\right), \\ &y^T\boldsymbol{\alpha}=0, \\ &0\le\boldsymbol{\alpha}_i\le C, i=1,2,...,n \end{aligned} \tag{4.5}$$

where e is a vector of all ones, C > 0 is the upper bound, Q is an n by n positive semi-definite matrix, $Q_{ij}=y_i y_j K\left(\mathbf{x}_i,\mathbf{x}_j\right)$, and $K\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left(\gamma\mathbf{x}_i^T\mathbf{y}_j+\zeta\right)^d$ is the kernel function. The decision function is:

$$f\left(\mathbf{x}\right)=\sum_{i=1}^{n} y_i\boldsymbol{\alpha}_i K\left(\mathbf{x}_i,\mathbf{x}\right)+b \tag{4.6}$$

The prediction of any test data x is $\operatorname{sign}\left(f\left(\mathbf{x}\right)\right)\in\left\{-1,1\right\}$.

For flare forecasting, the training and forecasting procedures of a support vector machine are as follows:

1. Training: The training data contain several samples; each sample is composed of four probabilities (the output of ordinal logistic regression) and one label (–1 or 1). If a given active region indeed belongs to one level, the label is 1; otherwise, the label is –1.
2. Forecasting: Given an active region, at first, figure out its three photospheric magnetic features. Then feed these three variables into the ordinal logistic model to generate the output which contains four probabilities. Finally, feed the four probabilities into the support vector machine trained above. If the output of the support vector machine is 1, the estimation is that the given active region belongs to

one level; otherwise, it does not.

4.4 Experimental Results

The proposed flare forecasting method is implemented in MATLAB [96], which contains a procedure to fit a logistic regression model. The implementation also utilizes LIBSVM [97], which is a software package for support vector classification. The parameters adopted for LIBSVM are as follows: nu-Support Vector Classification of polynomial kernel $K\left(\mathbf{x}_i, \mathbf{x}_j\right) \equiv \left(0.01\mathbf{x}_i^T \mathbf{y}_j\right)^3$.

Four different trained SVM classifiers are used to perform yes/no forecasting for four different levels. The outputs of the first step (four probabilities) and the corresponding labels are sent to the four SVM classifiers to train them in the second step. The training procedures are almost the same for the four SVM classifiers except that different labels are used, i.e., when training a level-n SVM classifier, the four probabilities and a label which indicates whether the given sample belongs to level-n are fed into the SVM classifier, where n = 0, 1, 2 or 3. Alternatively, a multiclass SVM classifier can be used. In that way, only one multi-class SVM classifier is needed instead of four different binary SVM classifiers.

Leave-One-Out cross-validation is used to assess the prediction performance. For 230 samples, during each test case, 229 samples are used for training, and the remaining one is used for testing. If the predicted result is the same as the observation, it is positive; otherwise, it is negative. The process is repeated 230 times. Different samples are used for training and testing each time.

To assess the performance of the proposed method, seven measurements are used, which are correctness, true positive, true negative, weighted true rate, positive accuracy, negative accuracy, and weighted accuracy. All these seven measurements can be derived from the contingency table of the experiment. For a given contingency table like Table 4.2, the seven measurements are as follows:

1. correctness = (a + d)/(a + b + c + d);
2. true positive = a/(a + b);
3. true negative = d/(c + d);
4. weighted true rate = a/(a+b) ∗(a+c)/(a+b+c+d)+d/(c+d) ∗ (b+d)/(a+b+c+d);
5. positive accuracy = a/(a + c);
6. negative accuracy= d/(b + d);
7. weighted accuracy = a/(a+c) ∗ (a+c)/(a+b+c+d)+d/(b+d) ∗ (b+d)/(a+b+c+d).

Table 4.2 A Sample Contingency Table

	Observation Positive	Observation Negative
Forecasting Positive	a	b
Forecasting Negative	c	d

To compare the performance of the proposed method with the Logistic-Regression-based method [24] and SVM-based method [84], experiments are performed on the same dataset and the experimental results are illustrated in Figures 4.6, 4.7, 4.8 and 4.9. These four figures contain not only the contingency tables of each experiment, but also bar charts to illustrate the seven measures derived from contingency tables to help compare the performances of the three different flare forecasting methods. Please note, among the seven measures, positive accuracy is the most important measure in flare forecasting in that a miss (forecasting no flare, but flares occur) is worse than a false alarm (forecasting the occurrence of a flare, but it does not occur). The higher the values of positive accuracy, the fewer events are missed.

Figures 4.6, 4.7, 4.8 and 4.9 show the forecasting results for levels zero, one, two and three respectively, e.g., for level zero forecasting, all these 230 active regions in the dataset are classified into two groups according to whether they belong to level zero, and then the forecasting models are trained, and then tested.

Predicting the occurrence of X-class flares is the most important task of flare forecasting. As it can be seen from panel (a) in Figure 4.9, the Logistic-Regression-based method does not work well for forecasting X-class flares. Only 1 of the 34 X-class flares is forecasted correctly. At the same time, the SVM-based method and the proposed method can correctly forecast 7 of the 34 X-class flares, which is an improvement over the Logistic-Regression-based method. From Figure 4.8, it can also be noticed that the proposed method outperforms the other two methods on level two (M-class flares) forecasting.

The experimental results also show that the proposed flare forecasting method outperforms the SVM-based method on level one and level three forecasting. However, the proposed method is surpassed by the SVM-based method on level two forecasting, but the difference is very small. The performances of these two methods on level zero forecasting are almost the same.

	Observation Positive	Observation Negative
Forecasting Positive	52	28
Forecasting Negative	11	139

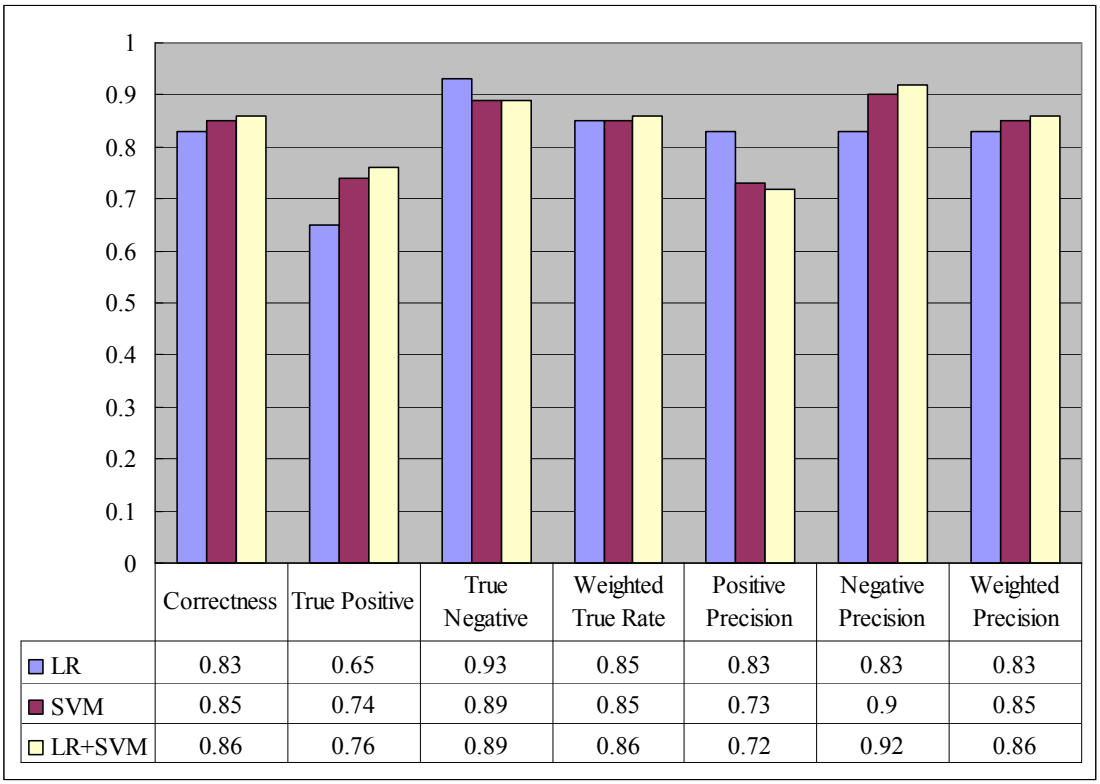
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	46	16
Forecasting Negative	17	151

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	45	14
Forecasting Negative	18	153

(c) Contingency table of proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Precision	Negative Precision	Weighted Precision
LR	0.83	0.65	0.93	0.85	0.83	0.83	0.83
SVM	0.85	0.74	0.89	0.85	0.73	0.9	0.85
LR+SVM	0.86	0.76	0.89	0.86	0.72	0.92	0.86

(d) Comparison of methods

Figure 4.6 Experimental results on level zero.

	Observation Positive	Observation Negative
Forecasting Positive	17	7
Forecasting Negative	48	158

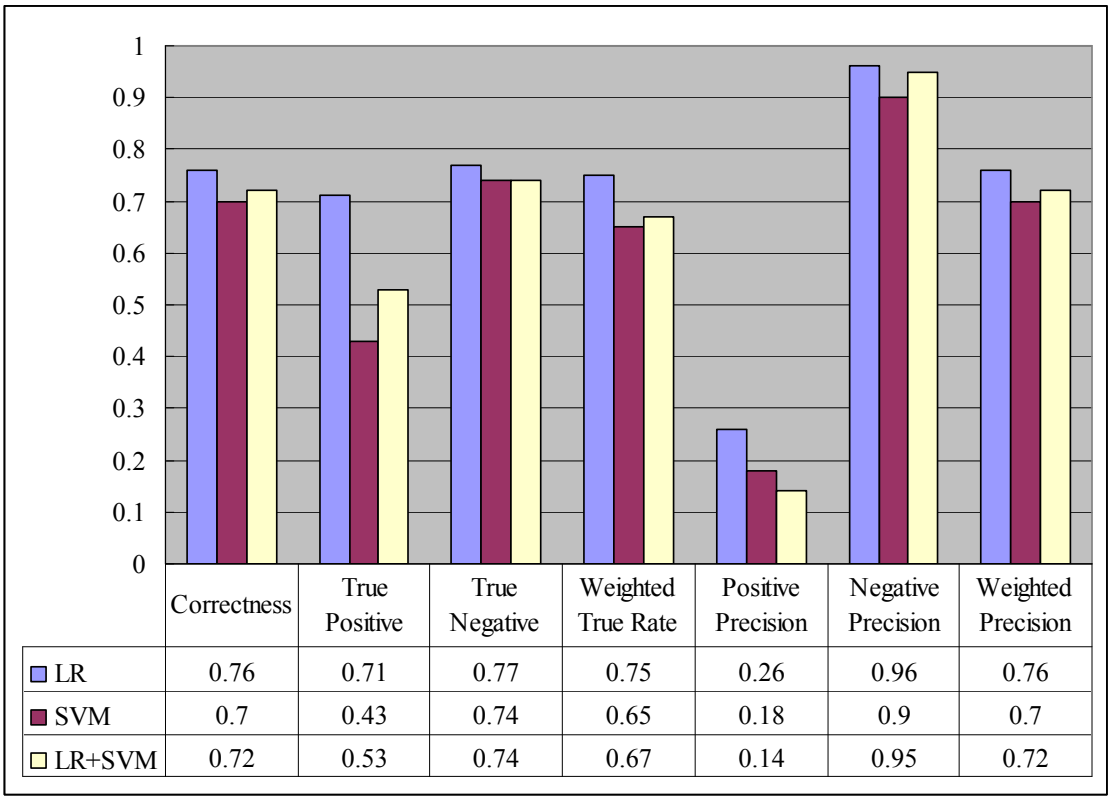
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	12	16
Forecasting Negative	53	149

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	9	8
Forecasting Negative	56	157

(c) Contingency table of proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Precision	Negative Precision	Weighted Precision
LR	0.76	0.71	0.77	0.75	0.26	0.96	0.76
SVM	0.7	0.43	0.74	0.65	0.18	0.9	0.7
LR+SVM	0.72	0.53	0.74	0.67	0.14	0.95	0.72

(d) Comparison of methods

Figure 4.7 Experimental results on level one.

	Observation Positive	Observation Negative
Forecasting Positive	10	2
Forecasting Negative	58	160

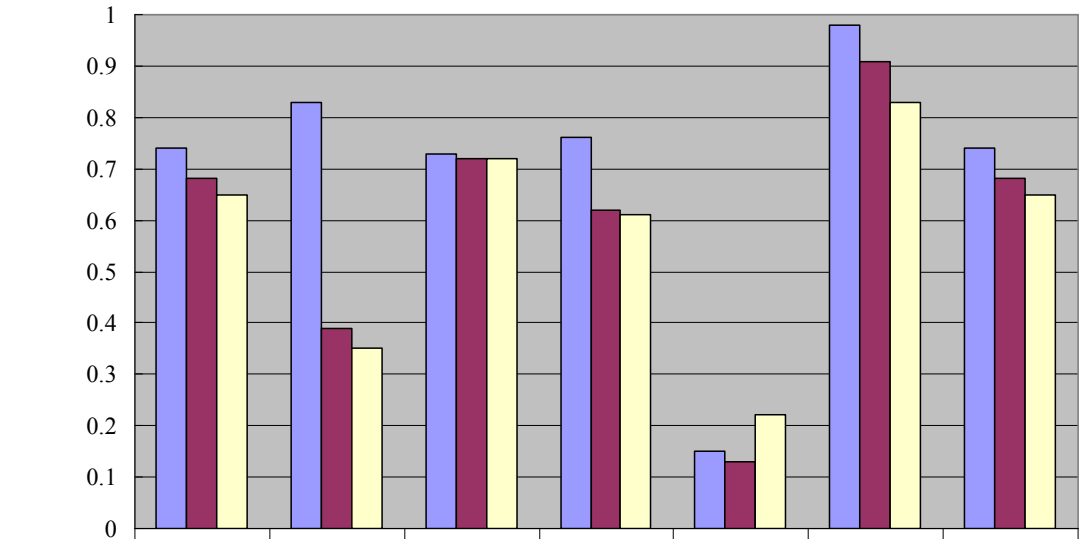
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	9	14
Forecasting Negative	59	148

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	15	27
Forecasting Negative	53	135

(c) Contingency table of proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Precision	Negative Precision	Weighted Precision
LR	0.74	0.83	0.73	0.76	0.15	0.98	0.74
SVM	0.68	0.39	0.72	0.62	0.13	0.91	0.68
LR+SVM	0.65	0.35	0.72	0.61	0.22	0.83	0.65

(d) Comparison of methods

Figure 4.8 Experimental results on level two.

	Observation Positive	Observation Negative
Forecasting Positive	1	0
Forecasting Negative	33	196

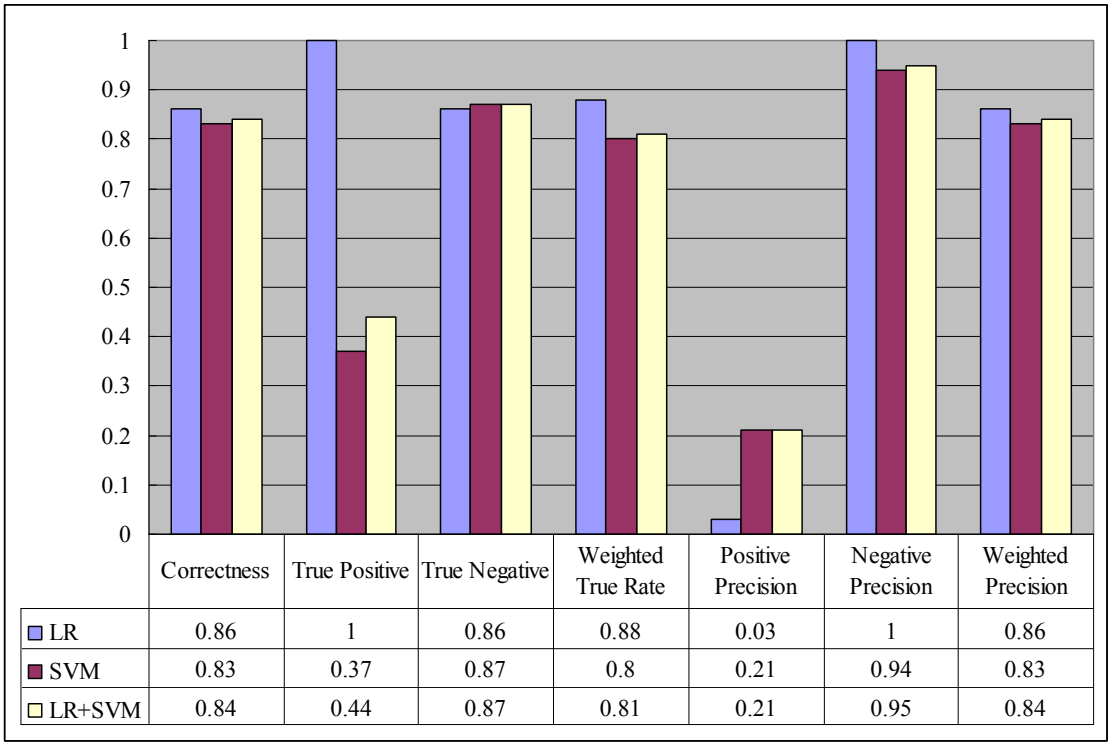
(a) Contingency table of logistic-regression-based method

	Observation Positive	Observation Negative
Forecasting Positive	7	12
Forecasting Negative	27	184

(b) Contingency table of SVM-based method

	Observation Positive	Observation Negative
Forecasting Positive	7	9
Forecasting Negative	27	187

(c) Contingency table of proposed method



	Correctness	True Positive	True Negative	Weighted True Rate	Positive Precision	Negative Precision	Weighted Precision
LR	0.86	1	0.86	0.88	0.03	1	0.86
SVM	0.83	0.37	0.87	0.8	0.21	0.94	0.83
LR+SVM	0.84	0.44	0.87	0.81	0.21	0.95	0.84

(d) Comparison of methods

Figure 4.9 Experimental results on level three.

4.5 Summary

In this chapter, a solar flare prediction method based on ordinal logistic regression and a support vector machine is proposed. For 230 active regions between 1996 and 2005, their magnetic parameters (L_{gpi}, T_{flux}, E_{diss}) are extracted from SOHO/MDI magnetograms and used for training. The experimental results can be summarized as follows:

1. The proposed method is a valid flare forecasting method, which performs almost equally well with the SVM-based method.
2. Although comparison shows that the positive accuracy of the proposed method is better than that of the Logistic-Regression-based method on X-class flare forecasting, the true positive rate (0.44) and positive accuracy (0.21) are still very low, meaning that it may fail to predict some occurrences of the X-class flares.
3. Since the proposed method is split into two cascading steps, one extra advantage of the proposed method over the SVM-based method is that it provide with confidence of the forecasting results. For example, when both of these two methods classify one active region into level three, it can derive the confidence level by examining the output of the first step. The output of the first step (the output of logistic regression) contains four probabilities (the four probabilities that a given active region belongs to the four levels). The higher the fourth probability, the more confidence can be put on the forecast results of X-class flares (corresponding to level three).

So far, the prediction model is limited to those magnetic parameters obtained only through SOHO/MDI magnetograms. There are several other physical parameters (such as magnetic free-energy, electric current and helicity injections) that can be used, and from

which it can be anticipated that the performance of the method can be improved. Similar to some other machine learning techniques, the proposed method is scalable with regard to accepting new parameters. In the future, after deriving several new magnetic parameters from vector magnetograms from the Solar Dynamic Observatory and *Hinode*, the new values should help to improve the performance of the proposed forecasting method. In addition, incorporating measures such as sunspot structure [98] and topology of solar magnetic fields [99] may also improve the performance of the proposed forecasting method.

CHAPTER 5

SOLAR FLARE FORECASTING USING SUNSPOT-GROUP CLASSIFICATION AND PHOTOSPHERIC MAGNETIC PARAMETERS

5.1 Introduction

Sunspots appear as dark spots on solar disks (illustrated in Figure 5.1) because they are cooler than its surroundings. Spots generally appear in pairs or groups, and thus astronomers classify them into different categories (sunspot-group classification). There are mainly two kinds of sunspot-group classification, namely McIntosh classification and Mount Wilson classification. McIntosh classification [100] is composed of a three-letter code which describes the class of sunspot group (single, pair and complex), penumbral development of the largest spot, and compactness of the group. The Mount Wilson classification [101] is used to describe the magnetic field structure. It seems that sunspot-groups, which are highly complex in appearance and magnetic, tend to give rise to solar flares [102].

Sunspot-group characteristics have long been used in solar flare forecasting and are still used extensively. Contarino *et al*. [103] studied sunspot-group parameters (i.e., Zrich class, magnetic configuration, area, morphology of the penumbra), and then performed a flare forecasting campaign based on the results. They claimed that the results obtained by comparing the flare forecasting probability with the number of flares that have actually occurred are quite encouraging. Kasper and Balasubramaniam [104] found that the penumbral area, umbral area and irradiance showed promise as possible parameters for predicting solar flares, particularly M-class flares. Qahwaji and Colak [86] compare the performances of several machine learning algorithm on flare forecasting using

classification of sunspot groups and solar cycle data. They found out that Support Vector Machines provide the best performance for predicting whether a classified sunspot group is going to flare.

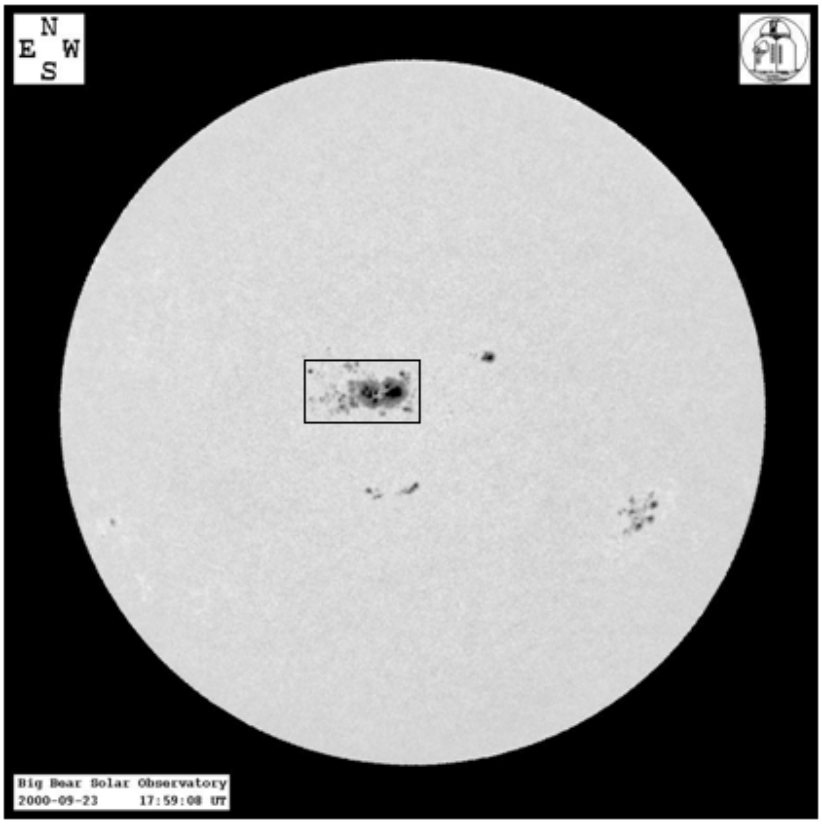


Figure 5.1 Illustration of a sunspot-group (enclosed in a black rectangle box) on Sep 23, 2000 (Courtesy of BBSO).

On the other hand, photospheric magnetic parameters derived from line-of-sight magnetograms are becoming more and more popular in solar flare forecasting. Jing *et al*. [20] studied the mean value of spatial magnetic gradients at strong-gradient magnetic neutral lines, the length of strong-gradient magnetic neutral lines and the total magnetic energy. They found that there exist statistical correlations between the three parameters of magnetic fields and the flare productivity of solar active regions. Yuan *et al*. [91] proposed

a cascading forecasting approach using total unsigned magnetic flux, length of the strong-gradient magnetic polarity inversion line, and total magnetic energy dissipation. Experimental results show that photospheric parameters are indeed can be used a precursor for solar flare forecasting.

In this study, aiming to improve the solar flare forecasting performance in the previous study conducted by Song *et al*. [24], both sunspot-groups classification and photospheric magnetic parameters are utilized. The solar flare forecasting is viewed as a classification problem in machine learning field. Given a testing sample, logistic regression is used to classify the sample to either a flaring sample or a non-flaring sample.

5.2 Dataset

To be consistent with the study conducted by Song *et al*. [24], the same dataset was used in this study. The dataset contains 230 samples from the year 1998 to 2005. Each sample is a pair of values describing the properties of an active region. A sample is composed of a label G_S indicating the classification of the sunspot-groups within the active region, a number T_{flux} indicating the total unsigned magnetic flux within the active region, a number L_{gnl} indicating the length of the strong gradient polarity neutral line and a label F indicating the level of the active region. According to the number and classes of flares produced by an active region, the level of an active region is defined as following: Level-0 if it produces no flares or only A-class and (or) B-class flares; Level-1 if it produces at least one C-class flare but no M- or X- class flares; Level-2 corresponds to those active regions which produce at least one M-class flare but no X-class flares; Level-3 corresponds to those active regions which produce at least one X class flare.

The classifications of the sunspot-groups G_S are extracted from the solar region summary [105] compiled by Space Weather Prediction Center of National Oceanic and Atmospheric Administration (NOAA). The magnetic parameters T_{flux} and L_{gnl} are derived from the magnetograms produced by the Michelson Doppler Imager (MDI), which is an instrument onboard the Solar and Heliospheric Observatory (SOHO). The flaring label F is derived from NOAA log of solar activity [106].

As mentioned in [24], the active regions used in the study were selected using the following criteria: (1) the center location of an active region is close to the solar disk center (within ±40 degrees in longitude and ±40 degrees in latitude); (2) the MDI full disk magnetograms are available; (3) since an active region may appear on the solar surface for a few days, it is treated as a different sample on different dates; (4) the magnetogram obtained at middle of each day by SOHO/MDI is chosen.

Table 5.1 illustrates a few samples from the constructed dataset, where sunspot-group classification is chosen from Mount Wilson Sunspot-group classification [101]. Totally, there are eight different classes of sunspot-groups. According to Taylor [101], the definition of Mount Wilson Sunspot-group classification is as follows:

Alpha: A unipolar sunspot group.

Beta: A sunspot group having both positive and negative magnetic polarities (bipolar), with a simple and distinct division between the polarities.

Gamma: A complex active region in which the positive and negative polarities are so irregularly distributed as to prevent classification as a bipolar group.

Beta-gamma: A sunspot group that is bipolar but which is sufficiently complex that no single, continuous line can be drawn between spots of opposite polarities.

Delta: A qualifier to magnetic classes (see below) indicating that umbrae separated by less than 2 degrees within one penumbra have opposite polarity.

Beta-Delta: A sunspot group of general beta magnetic classification but containing one (or more) delta spot(s).

Beta-Gamma-Delta: A sunspot group of beta-gamma magnetic classification but containing one (or more) delta spot(s).

Gamma-Delta: A sunspot group of gamma magnetic classification but containing one (or more) delta spot(s).

Table 5.1 A Few Samples from Dataset

Date	F	G_S	L_{gnl} $(403.0\ Mm)$	$T_{flux}\,(1.44\times10^{23}\ Mx)$
17/01/2005	0	Beta	0	0.0083
04/11/1998	1	Beta-Gamma	0.1687	0.2831
25/04/2001	2	Beta-Gamma-Delta	0.2333	0.7455
07/11/2004	3	Beta-Gamma-Delta	0.2184	0.2925

Total unsigned magnetic flux T_{flux} [20] and the length of the strong gradient polarity neutral line L_{gnl} [24] were defined in the previous chapter.

Figure 5.2 illustrates NOAA active region 0239 on Dec 31, 2002. Left panel shows the region itself. Middle panel shows the magnetic polarity inversion lines (blue lines) over-plotted on the smoothed region. Right panel shows the strong magnetic polarity inversion lines (blue lines) over-plotted on the smoothed region. To figure out magnetic polarity inversion line, the MDI magnetogram is firstly smoothed with a Gaussian filter with the standard deviation 10 and of size 30 by 30. And then contour lines at height zeros are find out (illustrated in middle panel). At last, the contour lines with strong gradient are

kept (illustrated in right panel). The length of the strong gradient magnetic polarity inversion line are figured out as L_{gnl}.

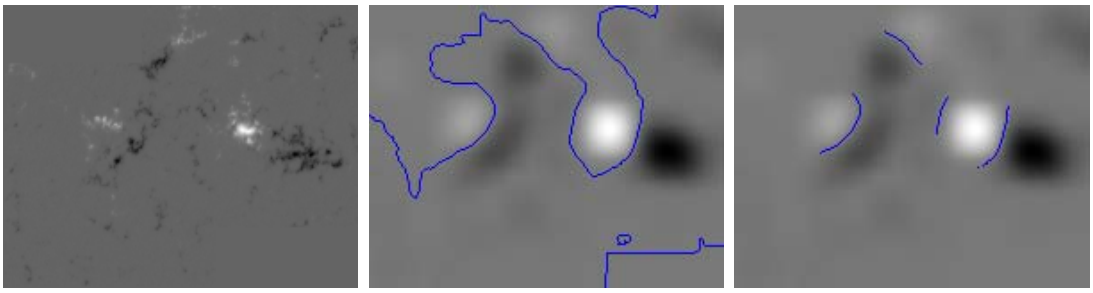


Figure 5.2 Illustrations of the active region NOAA 0239 on Dec. 31, 2002.

Figure 5.3 to Figure 5.6 illustrate scatter plots of some samples in our dataset grouped according to the sunspot-group classification. These scatter plots illustrate that the rate of flaring is different for samples belonging to different sunspot-group. It can be seen that samples in some sunspot-groups are more likely to produce strong flares. For example, Figure 5.3 contains data samples of Alpha sunspot-group. All those data samples are level-0 samples. Most data samples are level-2 and level-3 samples in Figure 5.6, which contains data samples of Beta-Gamma-Delta sunspot-group. The phenomena indicate that sunspot-group classification indeed provides another distinctive character of a data sample for flare-forecasting. This additional information may help us improve the flare forecasting performance.

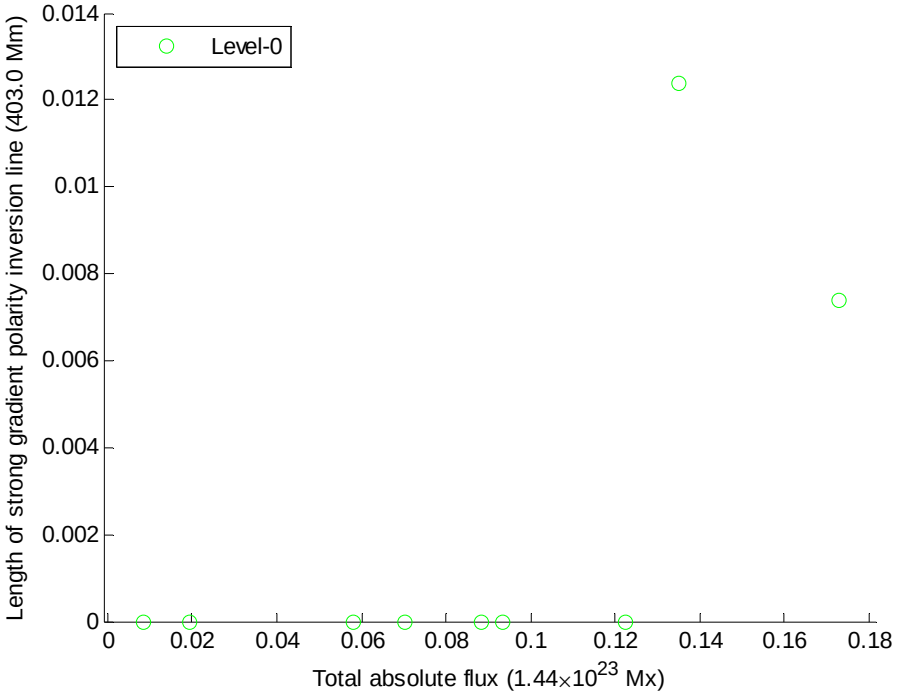


Figure 5.3 Scatter plot of data samples of Alpha sunspot-group.

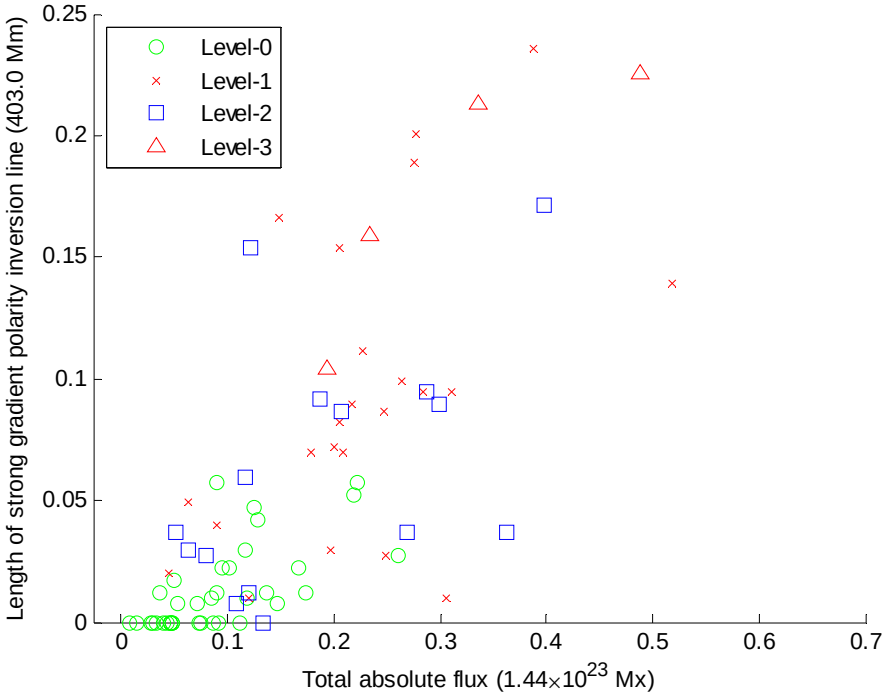


Figure 5.4 Scatter plot of data samples of Beta sunspot-group.

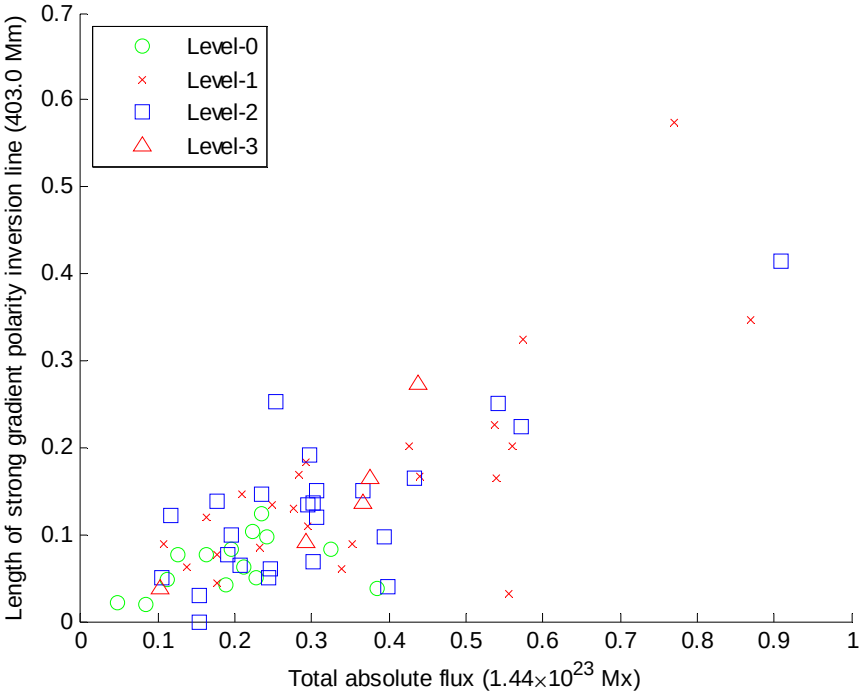


Figure 5.5 Scatter plot of data samples of Beta-Gamma sunspot-group.

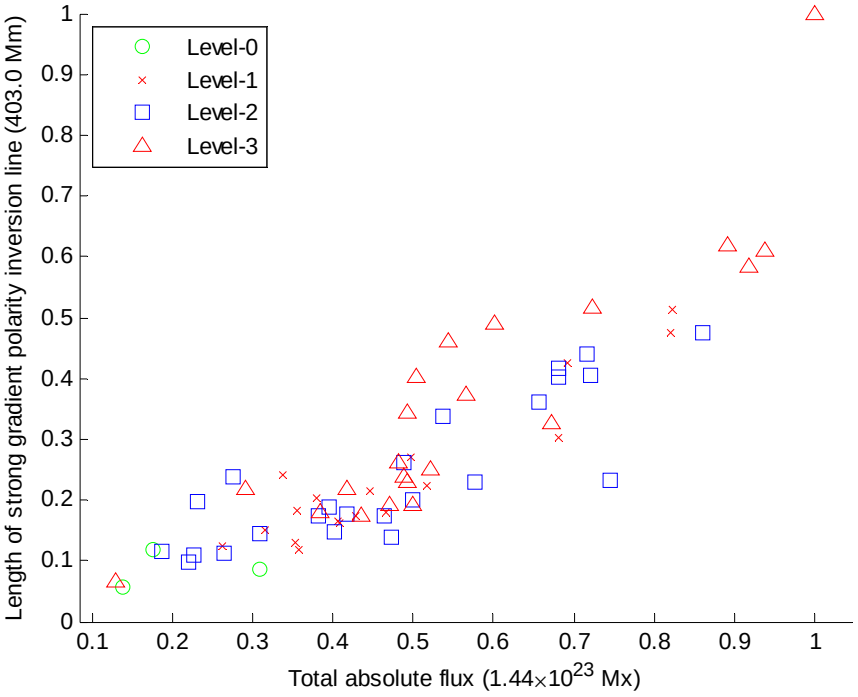


Figure 5.6 Scatter plot of data samples of Beta-Gamma-Delta sunspot-group.

5.3 Experimental Results

Logistic regression, as introduced in chapter one, is used to perform flare forecasting in this study. Three experiments were conducted. In the experiments, samples in the dataset were divided into two groups. In the first experiment, those samples which belong to level-0 were put into one group, and all other samples were put into another group. This experiment was designed to test the performance on predicting whether a given active region would produce class C-above (inclusive) flares. In the second experiment, those samples which belong to level-0 and level-1 were put into one group, and all other samples were put into another group. This experiment was designed to test the performance on predicting whether a given active region would produce class M-above (inclusive) flares. In the third experiment, those samples which belong to level-0, level-1 and level-2 were put into one group, and all other samples were put into another group. This experiment was designed to test the performance on predicting whether a given active region would produce X-class flares.

As introduced above, each data sample in the dataset contains three predictive parameters, namely, sunspot-groups classification G_S, total unsigned magnetic flux T_{flux} and length of strong gradient polarity neutral line L_{gnl}. In each experiment, seven tests are conducted to study the performance on different combinations of predictive parameters. The seven combinations are illustrated in Table 5.2. As shown in Table 5.2, each of first three combinations contains only one predictive parameter. The fourth combination to sixth combination each contains two predictive parameters. The seventh combination contains all the three predictive parameters.

Table 5.2 Seven Combinations of Predictive Parameters

Combination #	Predictive Parameters
1	T_{flux}
2	L_{gnl}
3	G_S
4	T_{flux}, L_{gnl}
5	$T_{flux}, G_S,$
6	$L_{gnl}, G_S,$
7	T_{flux}, L_{gnl}, G_S

Three metrics, namely accuracy, precision and recall [107], are used for evaluating the correctness of the flare forecasting method. Accuracy is a metric that computes the fraction of testing samples for which the forecasting are correct. Recall is computed as the fraction of correctly forecasted samples among all samples that actually produce flares, while precision is the fraction of correctly forecasted samples among those that the algorithm believes to belong to flaring samples. Recall can be seen as a measure of completeness, while precision is a measure of exactness.

$$\text{accuracy} = \frac{\begin{array}{c}\text{number of correctly forecasted flaring samples} + \\ \text{number of correctly forecasted non-flaring samples}\end{array}}{\text{total number of samples}} \tag{5.1}$$

$$\text{recall} = \frac{\text{number of correctly forecasted flaring samples}}{\text{total number of flaring samples}} \tag{5.2}$$

$$\text{precision} = \frac{\text{number of correctly forecasted flaring samples}}{\text{total number of forecasted flaring samples}} \tag{5.3}$$

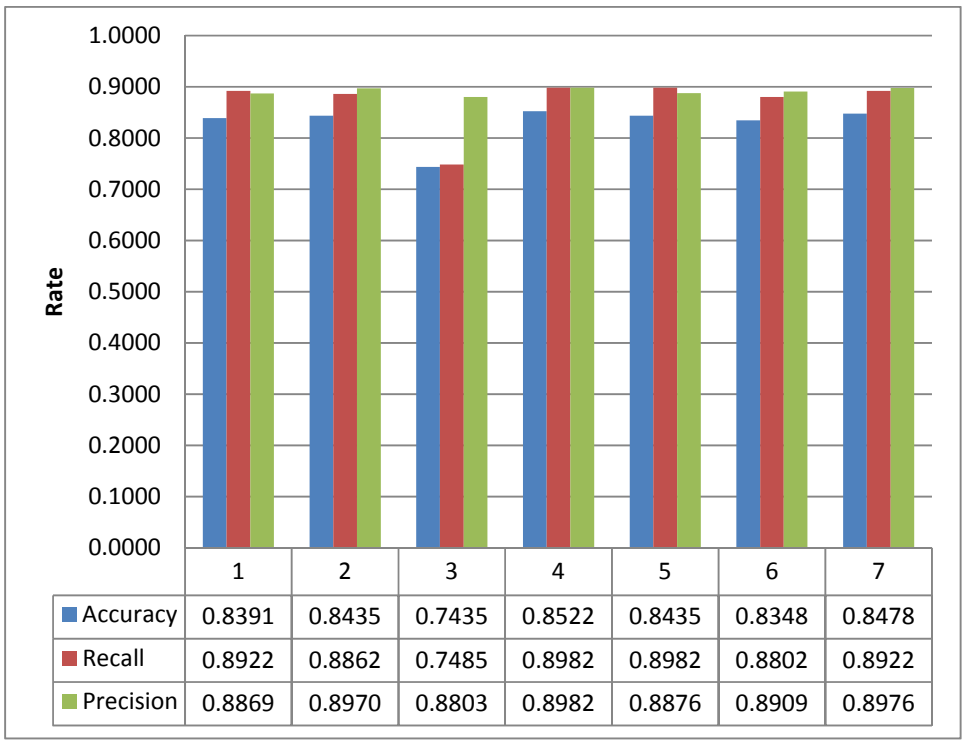
In settings where the goal of a machine learning method is prediction and one want to estimate how accurately a predictive model will perform in practice, cross-validation can be used [108, 109]. In this study, Leave-One-Out cross-validation [110, 111] is used for assessing how the proposed method will perform in practice. Because there are 230 samples in total, 230 iterations of training and testing need to be conducted. At each iteration, a distinct sample was chosen as a testing sample, the remaining 229 samples were used as training samples to train a logistic model. The testing sample was used to test the trained logistic model. Accuracy, precision and recall were figured out based on the results of 230 iterations.

Experimental results from the three experiments are illustrated in Figure 5.7, Figure 5.8 and Figure 5.9. For each combination of predictive parameters, three columns are drawn to represent the performance of flaring prediction measured by accuracy, recall and precision using color blue, red and green.

Figure 5.7 shows that the fourth combination (with predictive parameters T_{flux}, L_{gnl}) achieves the highest score measured with accuracy (0.8522), recall (0.8982) and precision (0.8982), while almost all other combinations perform quite well except the third combination (with predictive parameter sunspot-group classification G_S alone). The results show that the proposed method is very successful in class C-above (inclusive) flares forecasting in general.

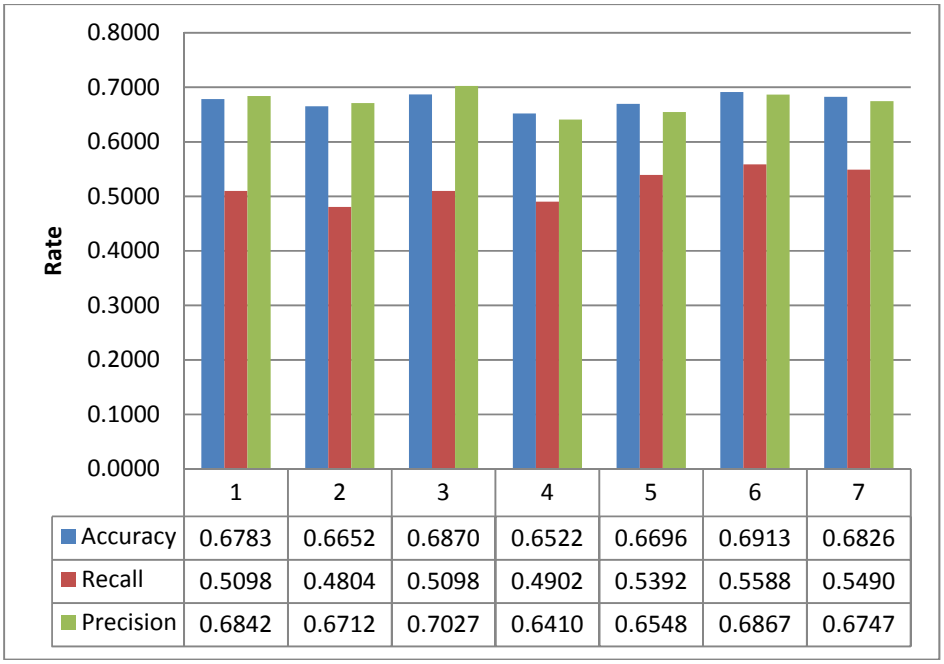
Figure 5.8 shows that the sixth combination (with predictive parameters L_{gnl}, G_S) achieves the highest score measured with accuracy (0.6913), recall (0.5588), while the third combination (with predictive parameter sunspot-group classification G_S alone) achieves the highest score measured with precision (0.7027).

Figure 5.9 shows that the fifth combination (with predictive parameters T_{flux}, G_S) achieves the highest score measured with accuracy (0.8565), precision (0.5714), while the fourth combination (with predictive parameters T_{flux}, L_{gnl}) and seventh combination (with predictive parameters T_{flux}, L_{gnl}, G_S) achieves the highest score measured with recall (0.2353). It is noted that, for the third combination (with predictive parameter sunspot-group classification G_S alone), the performance is very bad that the recall is zero and the precision cannot be derived because total number of forecasted flaring samples is zero.



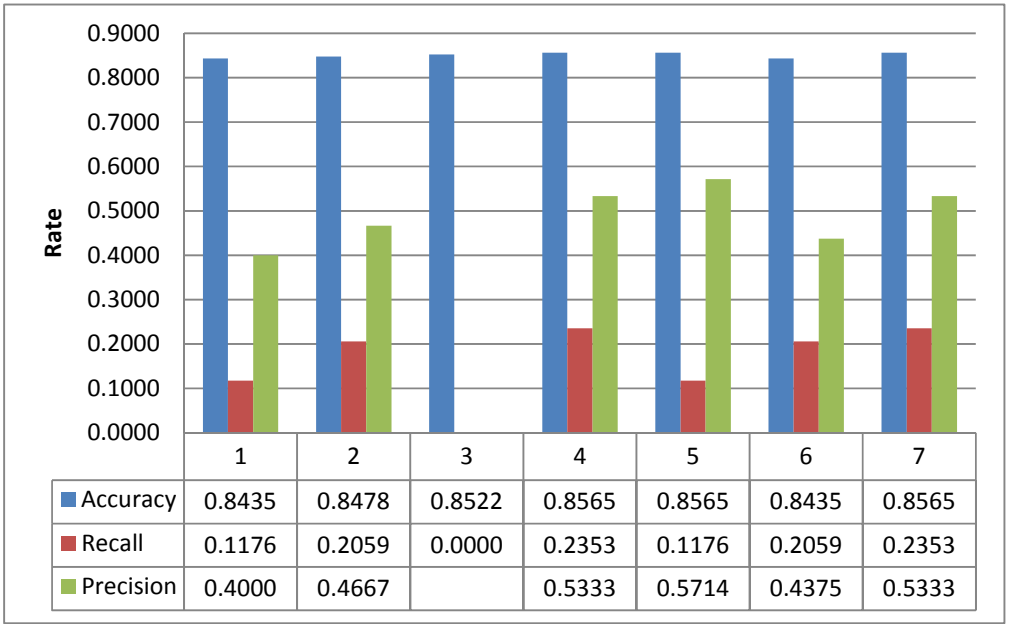
	1	2	3	4	5	6	7
Accuracy	0.8391	0.8435	0.7435	0.8522	0.8435	0.8348	0.8478
Recall	0.8922	0.8862	0.7485	0.8982	0.8982	0.8802	0.8922
Precision	0.8869	0.8970	0.8803	0.8982	0.8876	0.8909	0.8976

Figure 5.7 Accuracy, recall and precision of class C-above (inclusive) flare forecasting with seven different combinations of predictive parameters.



	1	2	3	4	5	6	7
Accuracy	0.6783	0.6652	0.6870	0.6522	0.6696	0.6913	0.6826
Recall	0.5098	0.4804	0.5098	0.4902	0.5392	0.5588	0.5490
Precision	0.6842	0.6712	0.7027	0.6410	0.6548	0.6867	0.6747

Figure 5.8 Accuracy, recall and precision of class M-above (inclusive) flare forecasting with seven different combinations of predictive parameters.



	1	2	3	4	5	6	7
Accuracy	0.8435	0.8478	0.8522	0.8565	0.8565	0.8435	0.8565
Recall	0.1176	0.2059	0.0000	0.2353	0.1176	0.2059	0.2353
Precision	0.4000	0.4667		0.5333	0.5714	0.4375	0.5333

Figure 5.9 Accuracy, recall and precision of X-class flare forecasting with seven different combinations of predictive parameters.

5.4 Summary

A flare forecasting method using both photospheric magnetic parameters and sunspot-group classification is presented in this chapter. Photospheric magnetic parameters are quantitive measurement of the magnetic fields of an active region. On the contrary, sunspot-groups classification is qualitative description of the magnetic configurations of the sunspot-group of an active region.

From the experimental results, it can be concluded that: (1) The overall performance of flare forecasting method is best on class C-above (inclusive) flare forecasting, but worst on X-class flare forecasting. (2) Sunspot-group classification alone is not a very good predictive parameter in flare forecasting. For X-class flare forecasting, it failed to make a single correct forecasting. (3) Although different combinations of the predictive parameters, except the combination that using sunspot-group classification alone, achieve similar scores in flare forecasting, the seventh combination (using predictive parameters T_{flux}, L_{gnl}, G_S) is most reliable in flare forecasting. (4) The presented method is not very applicable in X-class flare forecasting. Both recall and precision for X-class flare forecasting are very low, which means there will be false alarms and misses in practice.

As mentioned in the previous chapter, the prediction model is limited to those magnetic parameters obtained only through SOHO/MDI magnetograms. The key to predict solar flares lies in obtaining an accurate and complete picture of the structure of the magnetic field of the Sun [1, 16, 17]. In the future, after deriving several new magnetic parameters (such as magnetic free-energy [19], electric current [112] and helicity injections [18]) from vector magnetograms from the Solar Dynamic Observatory and

Hinode, the new values should help to improve the performance of the proposed forecasting method.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this dissertation, advanced algorithms are designed and implemented for solar filament detection, solar fibril tracing and solar flare prediction. These algorithms can be used for automatic processing of solar data to derive valuable information in the field of space weather research.

In filament detection, image enhancement, edge detection, segmentation, morphological operation and Hough transform are applied, which aims to automatically detect and characterize solar filaments in Hα solar images obtained from different solar observatories. Experimental results on 125 solar images captured by four different solar observatories show that the accuracy of the proposed method is more than 99% and 96% measured by area and by number of solar filaments, respectively. For filament characterization (such as heliographic centroid location), the center location and radius of the solar disks is identified using a two-stage Hough circle detection algorithm to mitigate the limitation imposed by traditional Hough circle detector. Experimental results show that the quality measure of the edge points obtained by median filter with Roberts edge operator can reach 74%. An application of the proposed filament segmentation method to filament tracking is illustrated. Preliminary results show the performance is very good. The area change of filaments are recorded, splitting and merging of filaments are also recorded, which presents a whole picture of the life span of filaments during a time period.

However, the proposed filament segmentation method is not very successful to detect filaments within active regions, where the brightness is much higher than average

brightness of solar disks. In the future, the method can be improved by adopting local thresholding and region growing.

A fibril tracing method is presented. Image processing techniques such as image enhancement, image segmentation, and union-find are used to segment fibrils from Hα images. Least squares curve fitting is used to model segmented fibrils. Experimental results show that the proposed method is very successful in segmentation and modeling of most fibrils, especially major fibrils. For future research, the least square fitting of fibrils can be improved by introducing optimization mechanism to search for a good balance between smoothness (low order polynomial fitting) and accuracy (high order polynomial fitting, but can cause oscillations).

In flare forecasting, logistic regression and support vector machine is used to predict solar flares based on properties of magnetic fields derived from SOHO/MDI magnetograms. To mitigate the limitations of logistic regression and support vector machine, a two-step prediction scheme is proposed which combines the forecasting probabilities of logistic regression and support vector machine. Experimental results illustrate that proposed method is a valid flare forecasting method, which performs almost equally well with the SVM-based method. Since the proposed method is split into two cascading steps, one extra advantage of the proposed method over the SVM-based method is that it provides confidence level of the forecasting results. It is also illustrated that the performance of flare forecasting can be improved by incorporation sunspot-group classification.

So far, the prediction model is limited to those magnetic parameters obtained only through SOHO/MDI magnetograms. There are several other physical parameters (such as

magnetic free-energy [19], electric current [112] and helicity injections [18]) that can be used, and from which it can be anticipated that the performance of the method can be improved. Similar to some other machine learning techniques, the proposed method is scalable with regard to accepting new parameters. In the future, after deriving several new magnetic parameters from vector magnetograms obtained by the Solar Dynamic Observatory and Hinode, the new values should help to improve the performance of the proposed forecasting method.

This dissertation presents several applications of computer science in solar physics. The proposed techniques, such as cascading Hough circle detector, adaptive image segmentation and statistical learning for prediction, can be applied to a broad range of fields. For example, the proposed techniques can be used to segment a region-of-interest from an X-ray computed tomography (CT), and predict the probability of the occurrence of a certain cancer based on some properties derived from the region-of-interest. The proposed techniques can also be used to trace the flow of the radioactive matters escaping from the Japan's Fukushima nuclear plant base on images obtained by sensors on satellites. In addition, the proposed techniques can be used to trace the flow of industrial sewage in contaminated rivers from images obtained by satellites.

The advancement and widespread of digital imaging techniques bring us images of higher resolution, quality and volume. It becomes very time-consuming to process those digital images. New technologies such as high performance computing (HPC) [113, 114], grid computing [115, 116] and cloud computing [117-119] can be utilized to significantly accelerate computing. In addition, GPU (graphics processing unit) computing [120, 121] is

rising in the field of digital image processing to help in vector manipulation. In the future, the proposed method can be modified and optimized to utilize those new technologies.

REFERENCES

1. Hanslmeier, A., The Sun and Space Weather. 2nd ed. Astrophysics and Space Science Library. Vol. 347. 2010: Springer. 315.
2. NRC, N.R.C., Earth Science and Applications from Space: National Imperatives for the Next Decade and Beyond. 2007: The National Academies Press.
3. Lang, K.R., The Cambridge Encyclopedia of the Sun. 2001: Cambridge University Press. 268.
4. Bhatnagar, A. and W. Livingston, Fundamentals of Solar Astronomy. World Scientific Series in Astronomy and Astrophysics. 2005: World Scientific Publishing Company.
5. Global Hydrogen Alpha Network. [cited 2011 March 1]; Available from: http://swrl.njit.edu/ghn_web/.
6. Gilbert, H.R., et al., Active and Eruptive Prominences and Their Relationship to Coronal Mass Ejections. The Astrophysical Journal, 2000. **537**: p. 503-515.
7. Gosling, J.T., et al., Geomagnetic activity associated with earth passage of interplanetary shock disturbances and coronal mass ejections. Journal of Geophysical Research, 1991. **96**: p. 7831-7839.
8. Jing, J., et al., On the Relation between Filament Eruptions, Flares, and Coronal Mass Ejections. The Astrophysical Journal,, 2004. **614**: p. 1054-1062.
9. Gao, J., H. Wang, and M. Zhou, Development of an Automatic Filament Disappearance Detection System. Solar Physics, 2002. **205**: p. 93-103.
10. Shih, F.Y. and A.J. Kowalski, Automatic Extraction of Filaments in Halpha Solar Images. Solar Physics, 2003. **218**: p. 99-122.
11. Bernasconi, P.N., D.M. Rust, and D. Hakim, Advanced Automated Solar Filament Detection And Characterization Code: Description, Performance, And Results. Solar Physics, 2005. **228**: p. 97-117.
12. Qu, M., et al., Automatic Solar Filament Detection Using Image Processing Techniques. Solar Physics, 2005. **228**: p. 119-135.
13. Socas-Navarro, J.d.l.C.R.a.H., Are solar chromospheric fibrils tracing the magnetic field? Astronomy and Astrophysics, 2011. **527**.
14. Severny, A., Solar Physics. 2004: University Press of the Pacific.
15. Phillips, K.J.H., Guide to the Sun. 1995: Cambridge University Press. 404.
16. Hill, S. and M. Carlowicz, The Sun. 2006: Harry N. Abrams, Inc. 240.
17. Mullan, D.J., Physics of the Sun: A First Course. 2009: Chapman & Hall. 390.
18. Park, S.-H., et al., Time Evolution of Coronal Magnetic Helicity in the Flaring Active Region NOAA 10930. The Astrophysical Journal, 2010. **720**(2): p. 1102-1107.

19. Jing, J., et al., Free Magnetic Energy and Flare Productivity of Active Regions. The Astrophysical Journal, 2010. **713**(1): p. 440-449.

20. Jing, J., et al., The Statistical Relationship between the Photospheric Magnetic Parameters and the Flare Productivity of Active Regions. The Astrophysical Journal, 2006. **644**(2): p. 1273-1277.

21. Georgoulis, M.K. and D.M. Rust, Quantitative Forecasting of Major Solar Flares. The Astrophysical Journal Letters, 2007. **661**(1).

22. Barnes, G., et al., Probabilistic forecasting of solar flares from vector magnetogram data. Space Weather, 2007. **5**.

23. Li, R., et al., Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting. Chinese Journal of Astronomy and Astrophysics, 2007. **7**(3): p. 441-447.

24. Song, H., et al., Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flare Predictions. Solar Physics, 2009. **254**(1): p. 101-125.

25. Pratt, W.K., Digital Image Processing: PIKS Scientific Inside. 4 ed. 2007: Wiley-Interscience.

26. Jensen, J.R., Introductory Digital Image Processing. 3 ed. 2004: Prentice Hall.

27. Dougherty, G., Digital Image Processing for Medical Applications. 2009: Cambridge University Press.

28. Burger, W. and M.J. Burge, Principles of Digital Image Processing: Fundamental Techniques 2009: Springer.

29. Gonzalez, R.C. and R.E. Woods, Digital Image Processing. 2007: Prentice Hall.

30. Yuan, Y., D. Huang, and D. Liu. An Integer Wavelet Based Multiple Logo-watermarking Scheme. in Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences. 2006. Hangzhou, Zhejiang, China: IEEE Computer Society.

31. Jean-Luc Starck , E.J.C., David L. Donoho, The curvelet transform for image denoising. IEEE Transactions on Image Processing, 2002. **11**(6): p. 670 - 684.

32. Ballard, D.H., Generalizing the Hough transform to detect arbitrary shapes. Pattern Recognition, 1981. **13**: p. 111-122.

33. Mitchell, T., Machine Learning. 1 ed. 1997: McGraw Hill Higher Education.

34. Marsland, S., Machine Learning: An Algorithmic Perspective. 1 ed. 2009: Chapman and Hall/CRC.

35. Ganapathiraju, A., J.E. Hamaker, and J. Picone, Applications of support vector machines to speech recognition. IEEE Transactions on Signal Processing, 2004. **52**(8): p. 2348 - 2355.

36. He, X., et al., Introduction to the Issue on Statistical Learning Methods for Speech and Language Processing. IEEE Journal of Selected Topics in Signal Processing, 2010. **4**(6): p. 913 - 916.

37. Gupta, N., et al., The AT&T spoken language understanding system. IEEE Transactions on Audio, Speech, and Language Processing, 2006. **14**(1): p. 213 - 222.

38. Kuhn, R. and R. De Mori, The application of semantic classification trees to natural language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995. **17**(5): p. 449 - 460.

39. Koha, L.H., Surendra Ranganath, and Y.V. Venkateshb, An integrated automatic face detection and recognition system. Pattern Recognition, 2002. **35**(6): p. 1259-1273.

40. Apolloni, B., et al., eds. Machine Learning and Robot Perception. 1 ed. Studies in Computational Intelligence. 2005, Springer.

41. Bishop, C.M., Pattern Recognition and Machine Learning. 1 ed. 2006: Springer.

42. Everitt, B.S., S. Landau, and M. Leese, Cluster Analysis. 4 ed. 2009: Wiley.

43. Hsu, C.-W., C.-C. Chang, and C.-J. Lin. A practical guide to SVM classification. 2010; Available from: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

44. Hosmer, D.W. and S. Lemeshow, Applied logistic regression. 2 ed. Wiley Series in probability and statistics. 2000: Wiley-Interscience Publication.

45. Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2009: Springer.

46. Kleinbaum, D.G. and M. Klein, Logistic Regression: A Self-Learning Text Statistics for Biology and Health. 2010: Springer.

47. Boser, B.E., I.M. Guyon, and V. Vapnik. A training algorithm for optimum margin classifiers. in Fifth Annual Workshop on Computational Learning Theory. 1992. Pittsburgh: ACM.

48. Cortes, C. and V. Vapnik, Support vector networks. Machine Learning, 1995. **20**: p. 1-25.

49. Minoux, M. and S. Vajda, Mathematical Programming: Theory and Algorithms. Wiley-Interscience series in discrete mathematics and optimization. 1986: John Wiley and Sons Ltd.

50. Gunn, S.R. Support Vector Machines for Classification and Regression. 1998 [cited March 22 2011].

51. Cristianini, N. and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. 1 ed. 2000: Cambridge University Press.

52. Schlkopf, B. and A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. 2001: The MIT Press.

53. Steinwart, I. and A. Christmann, Support Vector Machines. Information Science and Statistics. 2008: Springer.

54. Denker, C., et al., Synoptic Hα Full-Disk Observations of the Sun from Big Bear Solar Observatory – I. Instrumentation, Image Processing, Data Products, and First Results. Solar Physics, 1999. **184**(1): p. 87-102.

55. Jähne, B., H. Scharr, and S. Körkel, Handbook of Computer Vision and Applications. Academic Press. 1999.

56. Kimme, C., D.H. Ballard, and J. Sklansky, Finding circles by an array of accumulators. Commun. Assoc. Comput., 1975. **18**: p. 120-122.

57. Ozhogina, O.A., Solar limb darkening in the wings of the CaII H and K lines. Geomagnetism and Aeronomy, 2009. **49**(7): p. 879-883.

58. Thompson, W.T., Coordinate systems for solar image data. Astronomy and Astrophysics, 2006. **449**: p. 791 - 803.

59. Soille, P., Morphological Image Analysis: Principles and Applications. 2002: Springer.

60. Cormen, T.H., et al., Introduction to Algorithms. 2001: The MIT Press.

61. Filament Dataset. [cited 2011 March 17]; Available from: http://filament.njit.edu/dataset/.

62. Filament by hand. [cited 2011 March 17]; Available from: http://filament.njit.edu/hand/. .

63. Perreault, S. and P. Hebert, Median Filtering in Constant Time. IEEE Transactions on Image Processing, 2007. **16**: p. 2389 -2394.

64. Weiss, B., Fast median and bilateral filtering. ACM Trans. Graph., 2006. **25**: p. 519–526.

65. Canny, J., A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986. **8**: p. 679-698.

66. IDL [cited 2011 March 17]; Available from: http://www.ittvis.com/.

67. SolarSoftWare. [cited 2011 March 17]; Available from: http://www.lmsal.com/solarsoft/.

68. Beck, J.G., A comparison of differential rotation measurements. Solar Physics, 1999. **191**(1): p. 47-70.

69. Snodgrass, H.B. and R.K. Ulrich, Rotation of Doppler features in the solar photosphere. Astrophysical Journal, 1990. **351**: p. 309-316.

70. Casini, R., et al., Magnetic Maps of Prominences from Full Stokes Analysis of the He I D3 Line. The Astrophysical Journal Letters, 2003. **598**(1): p. 67.

71. Casini, R., R.M. Sainz, and B.C. Low, Polarimetric Diagnostics of Unresolved Chromospheric Magnetic Fields. The Astrophysical Journal Letters, 2009. **701**(1): p. 43.

72. Socas-Navarro, H., et al., Spinor: Visible and Infrared Spectro-Polarimetry at the National Solar Observatory. Solar Physics, 2006. **235**: p. 55-73.

73. Gary, D.E. and G.J. Hurford, Coronal temperature, density, and magnetic field maps of a solar active region using the Owens Valley Solar Array. The Astrophysical Journal, 1994. **420**(2): p. 903-912.

74. Tun, S.D., D.E. Gary, and M.K. Georgoulis, Three-dimensional Structure of a Solar Active Region from Spatially and Spectrally Resolved Microwave Observations. The Astrophysical Journal, 2011. **728**(1).

75. Sedgewick, R., Algorithms in Java: Parts 1-4. 3 ed. 2002: Addison Wesley.

76. Pratt, W.K., Digital Image Processing. 2001: John Wiley & Sons.

77. Hanson, C.L.L.R.J., Solving Least Squares Problems. Classics in Applied Mathematics. 1987: Society for Industrial Mathematics.

78. Dauphin, C., N. Vilmer, and A. Anastasiadis, Particle acceleration and radiation in flaring complex solar active regions modeled by cellular automata. Astronomy and Astrophysics, 2007. **468**(1): p. 273-288.

79. Qu, M., et al., Automatic Solar Flare Detection Using MLP, RBF, and SVM. Solar Physics, 2003. **217**(1): p. 157-172.

80. Qu, M., et al., Automatic Solar Flare Tracking Using Image-Processing Technique. Solar Physics, 2004. **222**(1): p. 137-149.

81. McIntosh, P.S., The classification of sunspot groups. Solar Physics, 1990. **125**: p. 251-267.

82. Gallagher, P.T., Y.-J. Moon, and H. Wang, Active-Region Monitoring and Flare Forecasting I. Data Processing and First Results. Solar Physics, 2002. **209**(1): p. 171-183.

83. G. Barnes, et al., Probabilistic forecasting of solar flares from vector magnetogram data. SPACE WEATHER, 2007. **5**.

84. Li, R., et al., Application of support vector machine combined with K-nearest neighbors in solar flare and solar proton events forecasting. Advances in Space Research, 2008. **42**(9): p. 1469-1474.

85. Georgoulis, M.K. and D.M. Rust, Quantitative Forecasting of Major Solar Flares. The Astrophysical Journal, 2007. **661**: p. 109-112.

86. Qahwaji, R. and T. Colak, Automatic Short-Term Solar Flare Prediction Using Machine Learning And Sunspot Associations. Solar Physics, 2007. **241**: p. 195-211.

87. Wheatland, M.S., A Bayesian Approach to Solar Flare Prediction. The Astrophysical Journal, 2004. **609**: p. 1134-1139.

88. Falconer, D.A., R.L. Moore, and G.A. Gary, A measure from line-of-sight magnetograms for predicting coronal mass ejections. Journal of Geophysical Research, 2003. **108**(A10).

89. Abramenko, V.I., et al., Signature of an Avalanche in Solar Flares as Measured by Photospheric Magnetic Fields. Astrophysics Journal, 2003. **597**(2): p. 1135-1144.

90. McCullagh, P. and J. Nelder, Generalized Linear Models. 2 ed. 1989: Chapman and Hall/CRC.

91. YUAN, Y., et al., Automated flare forecasting using a statistical learning technique. Research in Astronomy and Astrophysics, 2010. **10**(8): p. 785-796.

92. Kleinbaum, D.G. and M. Klein, Logistic Regression: A Self-Learning Text. 2 ed. 2002: Springer.

93. Hosmer, D.W. and S. Lemeshow, Applied logistic regression. 2000: Wiley-Interscience Publication.

94. Vapnik, V., The Nature of Statistical Learning Theory. 2 ed. 1999: Springer.

95. Fan, R.-E., P.-H. Chen, and C.-J. Lin, Working Set Selection Using Second Order Information for Training Support Vector Machines. The Journal of Machine Learning Research, 2005. **6**: p. 1889-1918.

96. Moler, C.B., Numerical Computing with Matlab. 2004: Society for Industrial Mathematics.

97. Chang, C.-C. and C.-J. Lin. LIBSVM: a library for support vector machines. 2001 [cited 2011 March 17]; Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

98. Chen, W.-Z., et al., A Statistical Study of Rapid Sunspot Structure Change Associated with Flares. Chinese Journal of Astronomy and Astrophysics, 2007. **7**(5): p. 733-742.

99. Zhao, H., et al., Determination of the Topology Skeleton of Magnetic Fields in a Solar Active Region. Chinese Journal of Astronomy and Astrophysics, 2008. **8**(2): p. 133-145.

100. McIntosh, P.S., The classification of sunspot groups. Solar Physics, 1989. **125**(2): p. 251-267.

101. Taylor, P.O., Observing the sun. Practical astronomy handbook series. 1991: Cambridge University Press.

102. Norquist, D.C., An Analysis of the Sunspot Groups and Flares of Solar Cycle 23. Solar Physics, 2010. **269**(1): p. 111-127.

103. Contarino, L., et al., Flare forecasting based on sunspot-groups characteristics. Acta Geophysica, 2009. **57**(1): p. 52-63.

104. Kasper, D. and K.S. Balasubramaniam. Sunspot Characteristics Associated with Solar Flares. in AAS Meeting #215. 2010: American Astronomical Society.

105. Solar region summary. [cited 2011 March 17]; Available from: http://www.swpc.noaa.gov/ftpdir/warehouse/.

106. NOAA log of solar activity. [cited 2011 March 17]; Available from: http://www.swpc.noaa.gov/ftpmenu/indices/events.html.

107. Manning, C.D., P. Raghavan, and H. Schütze, Introduction to Information Retrieval. 2008: Cambridge University Press.

108. Rodriguez, J.D., A. Perez, and J.A. Lozano, Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010. **32**(3): p. 569 - 575.

109. Cawley, G.C. and N.L.C. Talbot, Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. Neural Networks, 2004. **17**(10): p. 1467-1475.

110. Zollanvari, A., U.M. Braga-Neto, and E.R. Dougherty, On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers Pattern Recognition, 2009. **42**(11): p. 2705-2723.

111. Cawley, G.C. and N.L.C. Talbot, Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers Pattern Recognition, 2003. **36**(11): p. 2585-2592.

112. Owens, M.J., The Formation of Large-Scale Current Sheets within Magnetic Clouds. Solar Physics, 2009. **260**(1): p. 207-217.

113. Gabriel Mateescu, W.G., Calvin J. Ribbens, Hybrid Computing — Where HPC meets grid and Cloud Computing. Future Generation Computer Systems, 2011. **27**(5): p. 440-453.

114. Gruber, R., et al., High performance computing for partial differential equations Computers & Fluids, 2011. **43**(1): p. 68-73.

115. Iosup, A. and D. Epema, Grid Computing Workloads. IEEE Internet Computing, 2011. **15**(2): p. 19 - 26.

116. Moretti, C., et al., All-Pairs: An Abstraction for Data-Intensive Computing on Campus Grids. IEEE Transactions on Parallel and Distributed Systems, 2010. **21**(1): p. 33 - 46.

117. Stinchcombe, N., Cloud computing in the spotlight. Infosecurity, 2009. **6**(6): p. 30-33.

118. Pallis, G., Cloud Computing: The New Frontier of Internet Computing. IEEE Internet Computing, 2010. **14**(5): p. 70 - 73.

119. Rehr, J.J., et al., Scientific Computing in the Cloud. Computing in Science & Engineering 2010. **12**(3): p. 34 - 43.

120. Elble, J.M., N.V. Sahinidis, and P. Vouzis, GPU computing with Kaczmarz's and other iterative algorithms for linear systems Parallel Computing, 2010. **36**(5-6): p. 215-231.

121. Chang, J.Y., et al., GPU-friendly multi-view stereo reconstruction using surfel representation and graph cuts. Computer Vision and Image Understanding, 2011. **115**(5): p. 620-634.