# Automated Sunspot Extraction, Analysis and Classification

Varun Jewalikar
*Final Year, B.E. , Computer Engineering*
*Delhi College of Engineering, New Delhi*
*India. auv@dce.edu*

Siddharth Singh
*Final Year, B.E. , Computer Engineering*
*Delhi College of Engineering, New Delhi*
*India. dce.siddharth@gmail.com*

***Abstract—*** *This paper provides a generic automated approach to sunspot extraction, analysis and McIntosh classification from MDI Continuum and Magneto-gram images. Solar flares, cyclical tendencies of sunspot behavior and many important solar characteristics and their influences can be studied by sunspot analysis on large amounts of data over an elongated period of time, as well as a single quantum of data. This paper aims to provide a fast, efficient and a more accurate approach for this purpose. It also aims to provide a rigorous scheme for McIntosh Classification.*

***Keywords—*** Sunspots, Wolf Number, McIntosh Classification, MDI Continuum/Magneto-gram images.

## I. INTRODUCTION

Sunspots are features of the solar surface, arising due to intense magnetic activity and are observed as visibly dark regions compared to surrounding regions. Sunspot behavior has direct correlation with solar radiation and the solar climate. Also the magnetic variations caused due to sunspots affect the conditions of satellite communication. Sunspot groups when classified according to McIntosh Classification [2] are used for Solar Flare prediction. Due to these factors the study and prediction of sunspot behavior has become a very important facet of solar astronomy. This task was carried out manually for a long time and with increasing solar data due to various new solar missions and observatories, it becomes important to devise an efficient and automated technique for this task. Moreover this task is highly subjective and there is no unanimity between classification from different observers and an accurate automated procedure would solve this problem.

Sunspot data archiving has been carried out manually over the past 400 years and more recently, with the advent of World Wide Web, sunspot data in different formats has become freely available. Our approach works in a manner so as to not depend on the source size, resolution and bit depth of the image.

Important features which can be extracted from a solar image using this approach are the Wolf's number, McIntosh class of a sunspot group, sunspot cycle period and attributes of individual sunspots as well as sunspot groups like polarity, area, symmetry, group length etc.

The Wolf's Number/Sunspot Number (relative sunspot number) is an index of the number of sunspots and sunspot groups present on the sun.

As this approach will mark out each individual group and sunspot it would be very easy to quantify the Sunspot number. More information about a sunspot like the magnetic polarity will be extracted using a magneto-gram along with photographic data.

Our approach consists of three different phases:

1. Removal of unnecessary information from source images and preparation of image for later stages.
2. Intelligent Thresholding to separate sunspots from the background, detection of sunspots and clustering the sunspots into groups.
3. Classification of sunspot groups into various classes according to McIntosh scheme using modified Zurich classification.

The morphological approaches used earlier [1] for sunspot detection were not preserving the original shape and size of each sunspot found on the image, which our algorithm preserves. Final results of our testing suggest a significant run-time improvement. [3]

## II. DATA DESCRIPTION AND PRE-PROCESSING

This stage of the analysis deals with pre-processing of data and removal of singular features, which depend on the source. At the end of this phase the data would become independent of its source.

The testing of this algorithm has been done using SOHO MDI intensity-gram (Fig 1) and SOHO MDI magneto-gram images [5]. The MDI is basically the Michelson Doppler Imager onboard the Solar and Heliospheric Observatory. It generates intensity-grams and magneto-grams of the sun which are known MDI Continuum and MDI magneto-gram respectively.

These will be used as an example for the pre-processing phase. Similar techniques can be extended to data from different sources.

This preparation and pre- processing has the following stages:

1. Removal of unnecessary channels of a multi channeled image (gray scaling of intensity-gram data).

2. Removal of textual and other irrelevant information from the image outside the radii of the sun (timestamp, title, etc). For our implementation the radii of the sun was calculated using a simple Hough transform to find the largest circle in an image and all the data outside this radius was removed. (Fig 2)
3. The intensity-gram and magneto-gram are generally out of sync i.e. they are captured from different sources at different times and they might be of a different scale. Here we need to scale down and sync these images so that they correspond to exactly the same position of the sun.
4. The sunspot grouping algorithm and some of the sunspot classification criterion require a 3-dimensional model of the solar surface. To convert the 2-dimensional image information to heliographic coordinates, we use a conformal and explicit formula for stereographic projection:

$$(x, y, z) = \left(\frac{2X}{1+X^2+Y^2}, \frac{2Y}{1+X^2+Y^2}, \frac{-1+X^2+Y^2}{1+X^2+Y^2}\right).$$

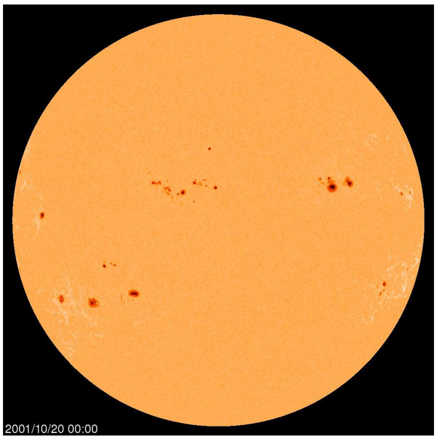

Fig 1 Original Image

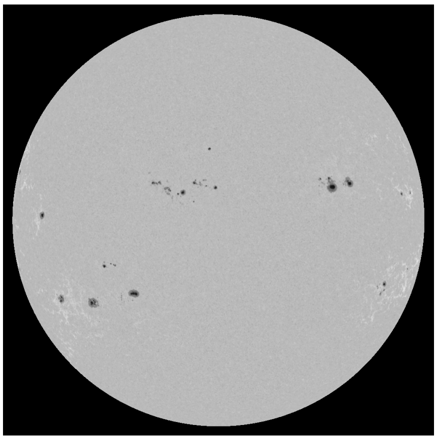
Fig 2 Image after application of stages 1 and 2

Data obtained from other sources might have features like limb darkening and these should be removed using approaches suggested here [1]. Other singular features depending on the source of the image are the exposure time and sensitivity of the sensor used to capture the image. The data should be rendered independent of these features by using suitable methods.

## III. THRESHOLDING AND CLUSTERING

### A. *Intelligent Thresholding of processed data*

The image data now has to be transformed to minimum required data-set with least errors possible. We start this by using an Intelligent Thresholding method. For this we are considering the classical sunspot definition, that any area darker than its surroundings is a sunspot. Here, we also use an empirical result which was tested on more than 100 images beforehand. Through an automated experiment we tried to find the minimum area of a sunspot on many images of different sizes and formats. The minimum sunspot area was recorded for each image and the final result obtained was that all the sunspots on all the images were a minimum of 0.5 heliographic degrees ($S_{min}$) in east-west span. We use this result as a sunspot definition along with the original with a 10% error margin.

Now that the pixel intensity values have been scaled down to 8-bit (1 channel) from 24-bit (3 channels) we have a range of 0 to 255 values for each pixel in our data set. Thresholding the whole image with a constant 'c' would replace all the pixels with values less than or equal to 'c' with a constant and the remaining values with another constant. The image thus becomes a binary image. Because our source image is free of any aberration and limb darkening error, unlike the method in [1], we can make the threshold method globally constant.

To find the optimal threshold for each image our algorithm uses a Binary Search combined with Connected Component Labeling algorithm. The search space in this case is the 256 valued range of all possible thresholds, and the optimal case is one where the east-west span of the smallest sunspot reaches the empirical constant($S_{min}$) found in the experiment mentioned earlier.

Because the threshold for different images can be highly varying as the image capturing method might vary from observatory to observatory, we use this dynamic thresholding method.

The steps in our algorithm are as follows:

1. Start a Binary state space search over the 256 possible different binary images obtained after applying different thresholds, with lower limit equal to 0 and upper limit equal to 255.
2. Find all the sunspots in the image under consideration using a 2 Pass Connected Component Labeling algorithm. Find the smallest area sunspot. Find its east-west span for the current threshold in its respective binary image.
3. If the value is greater than $S_{min}$ +0.05 then shift the lower limit of binary search to the current value, if it is less than $S_{min}$ +0.05 or greater than $S_{min}$ -0.05 then stop and return the current value, else shift the upper limit of binary search to the current value.
4. Continue till optimal value is found or the end of the algorithm is reached after log256 (=8) iterations.

$C \leftarrow \emptyset$
$T_{lower} \leftarrow 0$
$T_{upper} \leftarrow$ *palette range*
*while*$(T_{lower} < T_{upper})$ *do*
  $x \leftarrow 0$
  *while*$(x<x_{max})$
    $y \leftarrow 0$
    *while*$(y<y_{max})$
      *find (new component)*
      $C_{temp} \leftarrow$ *new component*
      $C \leftarrow C \cup C_{temp}$
    *end while*
  *end while*
  $S_{temp} \leftarrow span(min\_area(C))$
  *if* $(S_{temp} < S_{min}+0.5 \ \& \ min(s)>S_{min}-0.5)$
    *return* $(T_{lower} + T_{upper})/2$
  *else if* $(S_{temp} < S_{min}-0.5)$
    $T_{upper} = (T_{lower} + T_{upper})/2$
  *else if* $(S_{temp} > S_{min}+0.5)$
    $T_{lower} = (T_{lower} + T_{upper})/2$
  *end if*
*end while*
*return* $(T_{lower} + T_{upper})/2$

Fig 3 Algorithm for finding the optimal threshold

$x_{max}$ is the x resolution of the image, $y_{max}$ is the y resolution of the image, palette range is 256 for 8 bit images, min_area() returns the component with minimum area, span() calculates the east-west span of a component.

The time complexity of the algorithm given, in Big O notation is $O(log(palette\ range)*(x_{max}*y_{max})*2)=O(x_{max}*y_{max})$.

This is a an improvement over existing morphological methods [1] for the similar task, also our algorithm is using a direct thresholding method instead of a top-hat transformation and thus preserving the original shape and size of a sunspot.

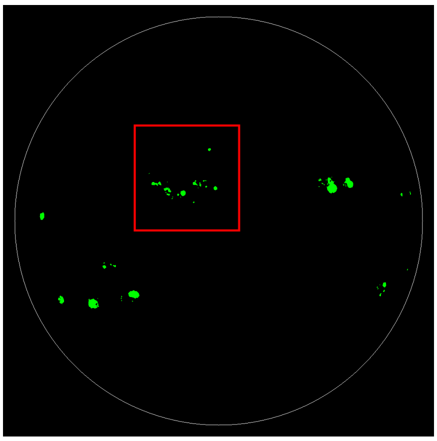

Fig 4 Image after thresholding, showing the area under consideration

### *B. Clustering of sunspots into groups*

Sunspots are observed to be formed in groups on sun's surface. These groups have their distinct physical properties and thus giving them characteristics to be analyzed and studied. Properties like polarity, group length, compactness, rudimentariness are what we try to extract from the group information. In order to find the group properties, the primary task is to identify each group on the surface. We are now using the heliographic coordinates for the images for exact superficial features analysis.

Various methods for finding sunspot groups are already present [4]. Most of these methods, although fairly accurate and methodical, are computationally expensive and take substantial time on large database of images. Our approach is focused on processing larger number of images and hence requires an algorithm which is faster and accurate still.

The empirical rule mentioned in [1] is what we use in our algorithm because of lower time complexity and ease of implementation, with being fairly accurate when compared to [4]. The rule states that a sunspot belongs to a group if the distance of the nearest sunspot belonging to that group is less than 6 heliographic degrees.

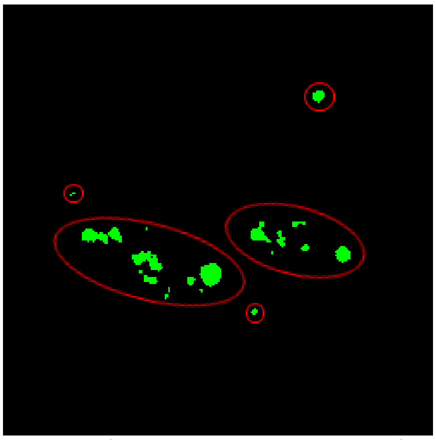

Fig 5 Image after clustering showing the groups found

## IV. McIntosh Classification of Sunspot Groups

Flare prediction from sunspot data relies heavily on their proper classification and these classes have to be such that they have a direct correlation with observed flares. Thus a simple system of sunspot classification is important for flare prediction. The two most widely used schemes for classification of sunspot groups are the Zurich classification scheme and the McIntosh classification scheme.

The Zurich classification scheme has been correlated with flares but even with its most active class (class F) the probability of a large flare is so low as to afford little real value to the forecast. Moreover aspects of sunspot groups, correlating with flares, have been observed, which are not a part of this scheme [2].

These flaws of the Zurich classification have led to the evolution of the McIntosh classification. In this scheme each sunspot group is represented with three letters as Zpc.

The three values are defined as:

Z-values: (Modified Zurich Sunspot Classification).
- A - A small single unipolar sunspot. Representing either the formative or final stage of evolution.
- B - Bipolar sunspot group with no penumbra on any of the spots.
- C - A bipolar sunspot group. One sunspot must have penumbra.
- D - A bipolar sunspot group with penumbra on both ends of the group. Longitudinal extent does not exceeds 10 deg.
- E - A bipolar sunspot group with penumbra on both ends. Longitudinal extent exceeds 10 deg. but not 15 deg.
- F - An elongated bipolar sunspot group with penumbra on both ends. Longitudinal extent of penumbra exceeds 15 deg.
- H - A unipolar sunspot group with penumbra.

p-values:
- x - no penumbra (group class is A or B)
- r - rudimentary penumbra partially surrounds the largest spot. This penumbra is incomplete, granular rather than filamentary, brighter than mature penumbra, and extends as litlle as 3 arcsec from the spot umbra. Rudimentary penumbra may be either in a stage of formation or dissolution.
- s - small, symmetric (like Zurich class J). Largest spot has mature,dark, filamentary penumbra of circular or elliptical shape with little irregularity to the border. The north-south diameter across the penumbra is less or equal than 2.5 degrees.
- a - small, asymmetric. Penumbra of the largest spot is irregular in outline and the multiple umbra within it are separated. The north-south diameter across the penumbra is less or equal than 2.5 degrees.
- h - large, symmetric (like Zurich class H). Same structure as type 's', but north-south diameter of penumbra is more than 2.5 degrees. Area, therefore, must be larger or equal than 250 millionths solar hemisphere.
- k - large, assymetric. Same structure as type 'a', but north-south diameter of penumbra is more than 2.5 degrees. Area, therefore, must be larger or equal than 250 millionths solar hemisphere.

c-values
- x - undefined for unipolar groups (class A and H)
- o - open. Few, if any, spots between leader and follower. Interior spots of very small size. Class E and F groups of 'open' category are equivalent to Zurich class G.
- i - intermediate. Numerous spots lie between the leading and following portions of the group, but none of them possesses mature penumbra.
- c - compact. The area between the leading and the following ends of the spot group is populated with many strong spots, with at least one interior spot possessing mature penumbra. The extreme case of compact distribution has the entire spot group enveloped in one continuous penumbral area.

PATRICK S. McINTOSH

TABLE I

Logic sequence for determining McIntosh sunspot types

Unipolar or bipolar?
Penumbra or no penumbra?
Penumbra on one end or both ends?
Length of group?
Rudimentary or mature penumbra?
Symmetric or asymmetric largest spot?
N–S diameter of largest spot?
Spots between leader and follower?
Mature penumbra in interior?

Fig 6 The logic sequence used as a reference from [2]

Based on the logic sequence it becomes clear that for correct classification several key attributes of a sunspot group mainly the polarity, the length, the features of the penumbra(if it is present), symmetry of the largest spot, the N-S diameter of the largest spot and the distribution of spots are needed.

But it is important to understand that not all of these features are needed for each spot for correct classification. For example there is no sunspot distribution for unipolar groups, there is only one type of largest spot for groups lacking penumbra and there is no compact distribution for classes B or C or groups whose largest spot has rudimentary penumbra.

Also the frequency of occurrence of each of these groups is not same. The short lived groups classes A and B occur most frequently among all while class E and F occur least frequently. Some combinations of parameters occur very rarely, they don’t even occur once during the entire sunspot cycle of 11 years [2].

On the basis of these factors we would like to suggest an approach which analyzes this classification problem as a learning heuristic state space search. It is important that each of the attributes of a sunspot group which leads to its correct classification be weighed differently i.e. only the most prominent features should be measured for each of the group. This state space search consist of 60 different nodes all of which occur with a different frequency, some of them occur all the time and some occur after very long spans of time. Hence the frequency of occurrence of each group over a very large period of time is taken as a guiding factor for this state space search.

The heuristic function H(n), can be defined as the probability of occurrence of that class over a large data set. Large data set signifies the number of days the sunspot groups have been observed for. The larger the data set the more accurately the heuristic function models the real life situation. This means that the data set spanning a large number of days and having at least a single occurrence of each of the 60 allowable classes should be chosen for highest accuracy.

For our implementation we chose a group of 200 images from various dates containing all types of distribution of sunspots and containing a variety of sunspot groups. We manually classified the sunspot groups in these images and obtained the probability distribution for each of the three features of the McIntosh classification (Zpc).

Taking any particular node (McIntosh Class) n, in this search space where, n=Zpc.

Then the Heuristic for this node n is given as:

*H(n)= Xz . Xp. Xc*

Xz = probability of Z in probability distribution (fig 1)

Xp = probability of p in probability distribution (fig 2)

Xc = probability of c in probability distribution (fig 3)

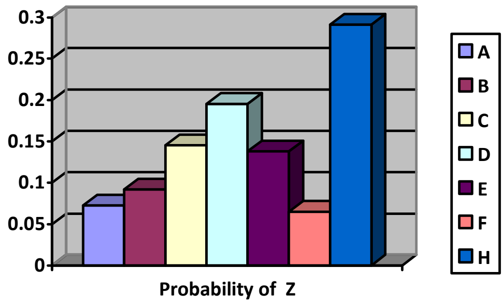


Fig 7

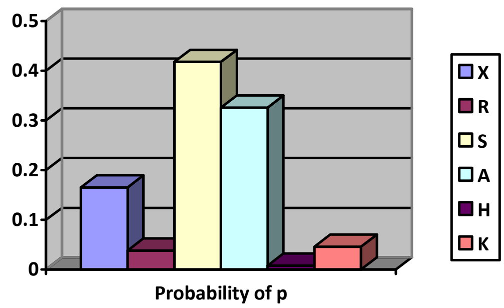


Fig 8

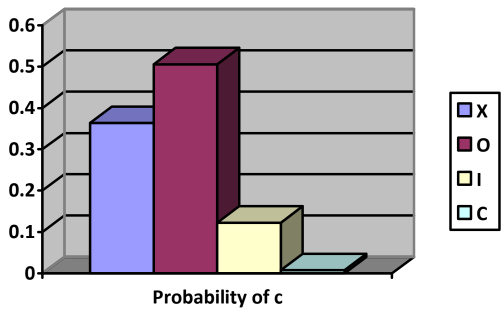


Fig 9

The search has a behavior of starting the search at the group (node) having the most probability of occurrence and going towards groups (nodes) which have decreasing probability. When it starts checking whether the current group under consideration belongs to the group specified by that node, it starts by checking the most distinguishing feature of that group first. This leads to a minimal time spent in verifying whether the current group belongs to that node or not. An image classified according to this scheme is shown in Fig 10.

After a successful classification of our group, the result is added back to the probability database making it a self-learning model. This is the first time this classification problem has been viewed from a heuristic based state space search. This algorithm has faster performance than previous classification approaches [3], [4] which give equal preference to each of the 60 allowable classes without taking the physical features of these classes into consideration.

## V. Implementation and Results

The above automated procedure was implemented and tested on a Core2duo 2.4 GHz 1Gb RAM PC running openSuse. The implementation was carried out using OpenCV library and testing was done on SOHO MDI intensitygram and magnetogram images.

The sunspot detection and grouping algorithm was tested on SOHO MDI archived images of the year 2001. This period was selected because it had high solar activity amongst all the periods for which archived images were available. The average Wolf's number was calculated for each month and it is plotted along with the manually tested data available at [5]. This sunspot detection and grouping algorithm has an accuracy of 94%. Fig 11 shows the plot.

The classification algorithm was tested on over 100 images which had a variety of group classes. These automatically classified images were then compared with manually classified images. Some of the images were correctly classified and in some images which have very dense sunspot groups the algorithm becomes a bit error prone. On average the accuracy of this algorithm is 65 per cent.

The whole process takes less than .1 sec on average on a single image. This shows a huge improvement over the approach developed in [3], which requires 4 sec on average for the same detection, grouping and classification algorithm.

## VI. Acknowledgment

We would like to thank Dr N. Rathnashree, Director of Nehru Planetarium, New Delhi for her invaluable guidance and support.

We would also like to thank Dr R.K. Sinha and Dr Rajni Jindal of Delhi Technological University for introducing us to the interesting fields of computer vision and artificial intelligence and listening to our ideas.

The images used for testing are courtesy of SOHO, a project of NASA-ESA.

## VII. References


[1] J. J. Curto, M. Blanca, E. Martínez, “Automatic Sunspots Detection on Full-Disk Solar Images using Mathematical Morphology”, Solar Physics,Vol 250, Issue 2, pp. 411-429, August 2008 .

[2] McIntosh, P.S. 1990, Sol. Phys., 125, 251.

[3] T. Colak, R. Qahwaji “Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images”, Solar Physics, Vol 248, Issue 2, pp. 277-296, April 2008.

[4] Nguyen, T. T., Willis, C. P., Paddon, D. J., Nguyen, S. H., and Nguyen, H. S. 2006. Learning Sunspot Classification. *Fundam. Inf.* 72, 1-3 (Apr. 2006), 295-309.

[5] The SOHO Archive, http://sohowww.nascom.nasa.gov/

[6] Dillencourt, M. B., Samet, H., and Tamminen, M. 1992. A general approach to connected-component labeling for arbitrary image representations. *J. ACM* 39, 2 (Apr. 1992), 253-280. DOI= http://doi.acm.org/10.1145/128749.128750

[7] K. J. H. Phillips. Guide to the Sun. Cambridge University Press, 1992.

[8] Meeus, J. 1998, Astronomical Algorithms - Second Edition, Willmann-Bell, Inc., Virginia.

[9] R. J. Bray and R. E. Loughhead. Sunspots. Dover Publications, New York, 1964.

[10] P. H. Scherrer, et al., Sol. Phys., 162, 129, 1995.

[11] Lanzerotti, L.J.: 2001, Space weather effects on communications. In: Daglis, I.A. (ed.) Space Storms and Space Weather Hazards, Kluwer, Dordrecht, 313 – 334.

[12] Zharkova, V., Ipson, S., Benkhalil, A., Zharkov, S.: 2005a, Feature recognition in solar images. Artif. Intell. Rev. 23, 209 – 266.

[13] Serra, J.: 1982, Image Analysis and Mathematical Morphology, Academic, New York.

[14] Zharkov, S., Zharkova, V., Ipson, S., Benkhalil, A.: 2005, Technique for automatic recognition of sunspots on full-disk solar images. EURASIP J. Appl. Signal Process. 15, 2573 – 2584.

[15] Dougherty, E.R.: 1992, An Introduction to Morphological Image Processing, SPIE Optical Engineering Press.

[16] Green, R.M.: 1985, Spherical Astronomy, Cambridge University Press, Cambridge.

[17] Bratsolis, E., Sigelle, M.: 1998, Solar image segmentation by use of mean field fast annealing. Astron. Astro-phys. 131, 371 – 375.

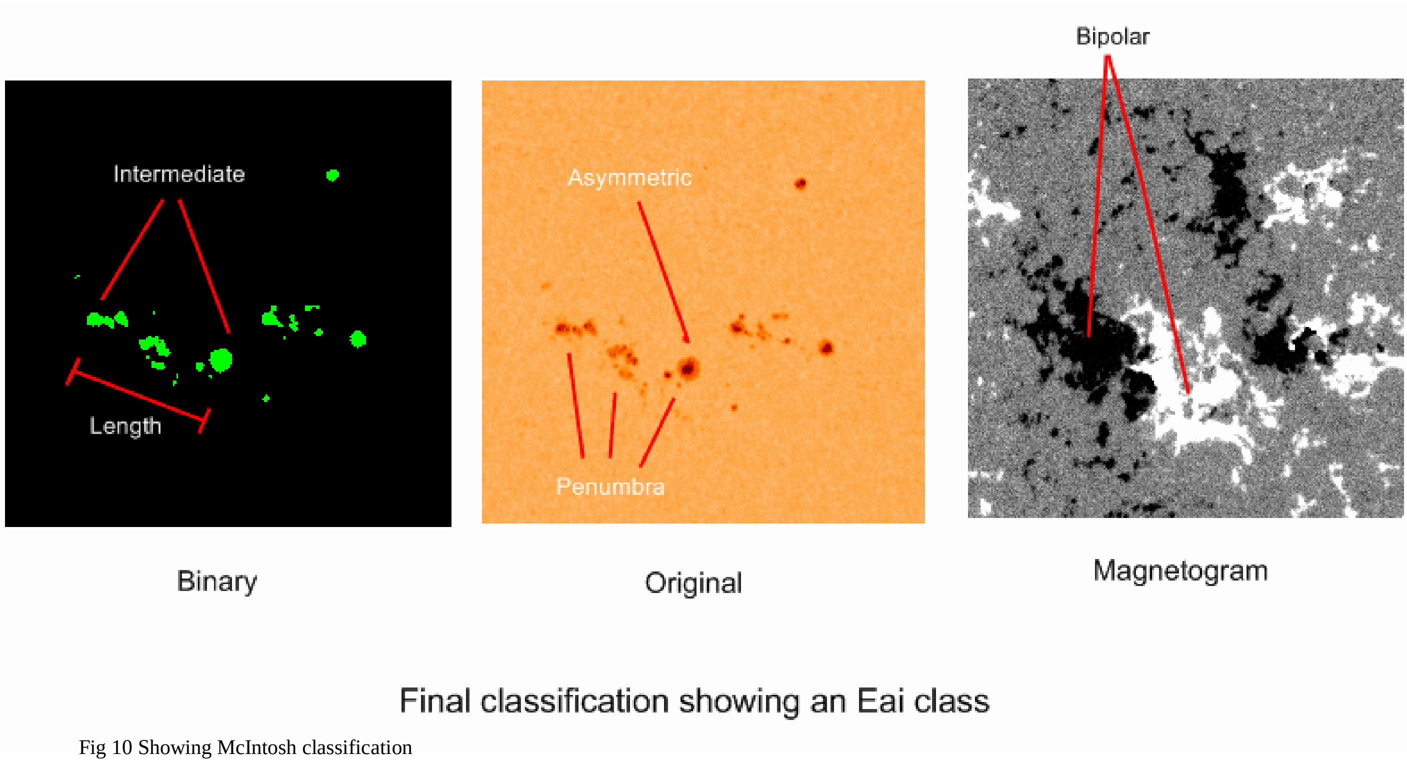


Fig 10 Showing McIntosh classification

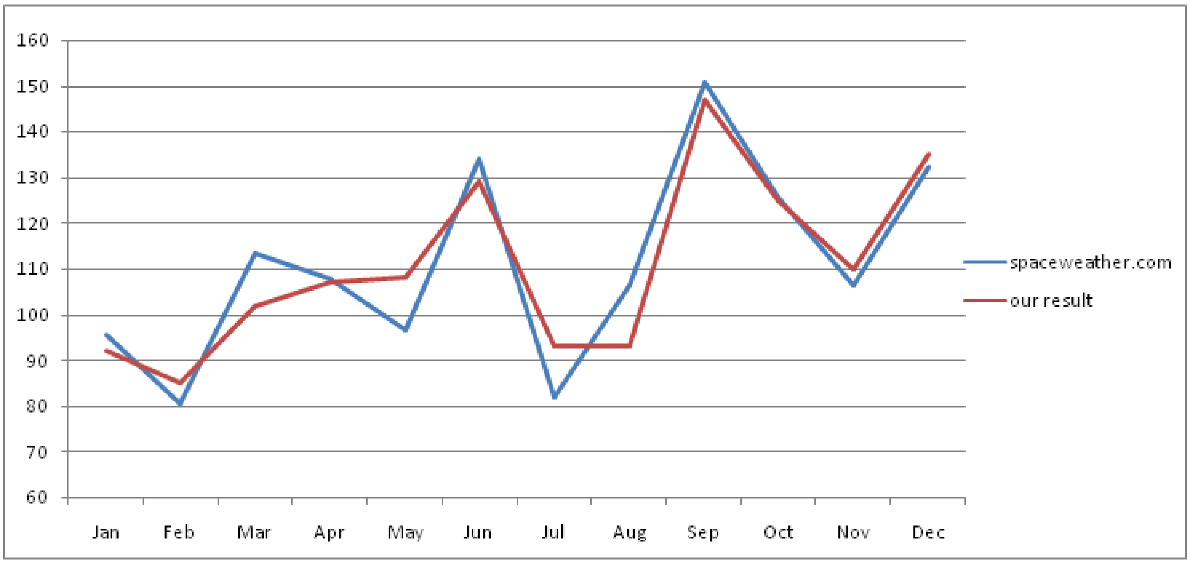


Fig 11 Showing the plot of sunspot number vs time