Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images

T. Colak and R. Qahwaji

Department of Electronic Imaging and Media Communications

University of Bradford, Richmond Road, Bradford BD7 1DP, England, UK

(E-mail: t.colak@bradford.ac.uk, r.s.r.qahwaji@bradford.ac.uk)

Abstract. A hybrid system for the automated detection and McIntosh-based classification of sunspot groups on SOHO/MDI white-light images using active-region data extracted from SOHO/MDI magnetogram images is presented in this paper. After sunspots are detected from MDI white-light images they are grouped/clustered using MDI magnetogram images. By integrating image-processing and Neural Networks techniques, detected sunspot regions are classified automatically according to the McIntosh classification system. Our results show that the automated grouping and classification of sunspots is possible with a high success rates when compared to the existing manually created catalogues. In addition our system can detect and classify sunspot groups in their early stages, which are usually missed by human observers.

1. Introduction

The observation, analysis, and classification of sunspots form an important part in furthering knowledge about the Sun, solar weather and its effect on Earth (Phillips, 1992). Previous research on solar flares showed that they are mostly related to sunspots and active regions (Künzel, 1960; Severny, 1965; Warwick, 1966; Sakurai, 1970; McIntosh, 1990). Sunspots are part of active regions, and their local behaviour is used for the forecast of solar activity (Hathaway, Wilson, and Reichmann,1994).

In this study, we present a computer platform for the automated detection, grouping and then classification of sunspots. In daily life, sunspot classification is mostly carried out manually by experts. This is a subjective, time consuming, and labour-intensive process and although classification rules are well defined, there is not always 100% unanimity in the resulting classification of sunspot groups between solar physicists even when working together. Accurate objective classification of sunspots can solve the unanimity problem faced by various solar observatories and space-weather-prediction groups around the world. Another argument supporting the use of such systems is the expected increase in solar data because of the new space missions.

Previous attempts at the detection of sunspots are reported in Curto, Blanca, and Solé (2003), Zharkov *et al.* (2004), and Nguyen, Nguyen, and Nguyen (2005). In Zharkov *et al.* (2004) an automated system for the detection of sunspots on the Ca K1 and Solar and Heliospheric Observatory (SOHO)/ Michelson Doppler Imager (MDI) white light images was presented and a detection rate of 98% was achieved for MDI images when compared with the Locarno Solar Observatory detection results. Nguyen, Nguyen, Nguyen (2005) used image processing, and clustering methods on SOHO/MDI white-light images for the recognition and classification of sunspots according to the modified Zurich class of the McIntosh system. Testing involved 128 sunspot groups. Although 100% correct classification rate was achieved for the modified Zurich classes C and H (25% of test data), only 60%, 19% and 21% correct classification rates were obtained for D, E and F (73.5% of the test data) were obtained respectively. Also, Curto, Blanca, and Solé (2003) used full disk white light images to automatically detect and cluster sunspots into groups. Sunspots were detected using morphological image processing techniques and neural networks were used to classify them. However, no good results were reported for grouping. Previous research shows that accurate detection of sunspots has been achieved on white light solar images. However, no good results were reported for the grouping and clustering of sunspots, which is the main reason behind the classification errors. This is the biggest challenge facing the creation of a fully automated and accurate sunspot classification system, as highlighted by Nguyen, Nguyen, Nguyen (2005) and Curto, Blanca, and Solé (2003).

In this work we present a system that uses SOHO/MDI intensitygram and magnetogram images to detect, group, cluster, and classify sunspots based on the McIntosh classification system. This is the first time, to our knowledge, that a computer platform is created to carry out this process automatically and objectively. Although MDI images are used in this work we believe that the principles and methods described here can be used by other researchers for processing different solar images with little modifications.

This paper is organized as follows: The types of images used are described in Section 2. The automated detection and grouping of the sunspots is introduced in Section 3. The classification of sunspot groups is described in Section 4, while the practical implementation and evaluation of the algorithms is reported in Section 5. Finally, the concluding remarks are given in Section 6.

2. Data Description

In this study, SOHO/MDI intensitygram images are used for sunspot detection, while SOHO/MDI magnetogram images are used for the detection of active regions. All of the images used are downloaded from the MDI website (http://soi.stanford.edu/) and they are in Graphics Interchange Format (GIF) format. Unlike the FITS images, GIF images do not contain a header file with observational information. Although using FITS images can decrease the error rate in the calculations and save processing times, we choose to use GIF images in order to combine this system with the automated flares prediction system described in Qahwaji and Colak (2007) in the very near future. This hybrid system will download online MDI continuum images in GIF format, detect sunspots, classify them and feed the classification results to the machine-learning system to predict whether a major flare is likely to occur in the short-term.

The MDI instrument on SOHO provides almost continuous observations of the Sun in the white-light continuum, in the vicinity of the Ni I 6767.8 Å photospheric absorption line. White-light pictures show how the Sun appears to the naked eye and MDI intensitygram images are primarily used for sunspot observations. The MDI data is available in several processed levels. The MDI images used in this research are level-2 images, which are smoothed, filtered, and rotated (Scherrer *et al.,* 1995). SOHO provides two to four MDI intensitygram images per day and twice as much magnetogram images with continuous coverage since 1995.

MDI magnetogram images help in measuring the magnetic field strengths on the Sun's photosphere. The magnetogram images show the magnetic fields of the solar photosphere, with black and white areas indicating opposite magnetic polarities. The dark areas are regions of South magnetic polarity (pointing toward the Sun) and the white regions have North magnetic polarity (pointing outward). These images can be used for detecting active regions. In daily life magnetogram images are used by observatories to decide and cluster sunspot groups. We believe that combining intensitygram images and magnetogram images will help us to decide and cluster sunspot groups in a similar way to the observatories.

3. Sunspot Detection and Grouping

Several stages are involved in the detection and grouping of sunspots, such as: pre-processing, initial detection of features (sunspots from intensitygrams, active regions from magnetograms), and clustering. All of these stages are described below.

3.1 Pre-processing of MDI Images

We divided pre-processing into two stages. The first stage is applied to intensitygram and magnetogram images and is called "*Stage-1*" processing. This stage involves detecting the solar disk, determining its centre and radius, calculating the solar coordinates, and filtering irrelevant information (*i.e.*, direction and date marks). "*Stage-2*" processing is applied to magnetogram images only and it is important because it enables us to correlate both MDI images. Usually there is a time difference (usually less than 30 minutes) between magnetogram and intensitygram images, and the size of the solar disk on both images could differ. The time difference problem has to be tackled in order to align these images and hence correlate them. To achieve this, magnetogram images need to be resized to have the same centre and radius as the intensitygrams, and their rotation across the solar disk corrected. This is very important because different magnetogram and white light images from different observatories can then be used for sunspot grouping and classification by applying the same conversion principle. These stages can be summarized as follows:

- *Stage-1*:
    - Apply the filtering process reported in Qahwaji and Colak (2006a, 2006b). Detect the solar disk, determine its radius and centre and create a mask.
    - Remove any information or marks (*i.e.,* date and direction) from the image using the mask created.
    - Calculate the Julian date by parsing the date and time information of the image from its name Meeus (1998).
    - Using the Julian date, calculate solar coordinates (The position angle, heliographic latitude, heliographic longitude) for the image using the equations in Meeus (1998). Although images are from the SOHO satellite, in this work, the solar coordinates (The position angle, heliographic latitude, and heliographic longitude) are calculated for the Earth view. We have carried out empirical studies and this will cause less than 1% error in our calculations which does not seem to have significant impact on the outcomes of this research.
- *Stage-2*:
    - Map the magnetogram image from Heliocentric-Cartesian coordinates to Carrington-Heliographic coordinates.
    - Re-map the image to Heliocentric-Cartesian coordinates. Use centre, radius, and solar coordinates of the intensitygram image as the new centre, radius, and solar coordinates of the magnetogram image.

Figure 1 shows the *Stage-2* processing example for a magnetogram image that was processed using *Stage-1* processing (Figure 1a), which is first mapped to the Heliographic coordinates (Figure 1b) and then re-mapped to the Heliocentric-

Cartesian coordinates using a new radius but the same solar coordinates (Figure 1c). Figure 1b, which is represented in heliographic coordinates, is shifted in this example for better view. The difference (Figure 1d) shows the data change, which is visible especially near the solar limb. This change is caused by the fact that the solar disk is remapped with a smaller radius.

Figure 2 shows the *Stage-2* processing example for two original magnetogram images marked as “2a” created on 27 July 2002 at 23:59 and “2b” created on 29 July 2002 at 01:35. *Stage-1* and *Stage-2* processing are applied to both images and the resulting images are shown as images 2c and 2d, respectively. Figure 2a is re-mapped to the Heliocentric-Cartesian coordinates with its previous solar coordinates and a new radius, while Figure 2b is re-mapped with the solar coordinates of Figure 2a and a new radius. The result of this time shift can be seen clearly in Figure 2.d, which has an information loss towards the West of the solar limb caused by the rotation of Sun during the 25 hour time difference.

3.2 Initial Detection of Solar Features

Initial detection of sunspots from intensitygram images and active regions from magnetogram images is carried out using intensity filtering and region growing methods, in a manner similar to Qahwaji and Colak (2006a).

The threshold value (**T_f**) for each image is found automatically using Equation (1), where, μ is the mean, σ represents the standard deviation, and α is a constant that is determined empirically based on the type features to be detected and images:

$$\mathrm{T_f} = \mu \pm (\sigma \times \alpha) \tag{1}$$

In order to detect sunspot candidates from intensitygram images a threshold value is calculated using Equation (1) with the minus (-) sign and 2.7 as the value of α. All of the solar disk pixels are compared with this threshold value. If the intensity value of the pixel is less than the threshold value, it is marked as a sunspot candidate.

Two threshold values have to be determined to detect the active region candidates in magnetogram images. The first threshold is used for detecting seeds with North magnetic polarity and the second is used for detecting seeds with South magnetic polarity. The value of the first threshold is determined using Equation (1) with a plus (+) sign and α equals two. All pixels that have intensity values larger than this threshold are marked as active region seeds with North polarity. In the same manner, the second threshold is determined using Equation (1) with the minus (-) sign and α equals two. Any pixel with an intensity value less than this threshold is marked as an active region seed with South polarity.

To find the optimum vale of α intensive experiments are carried out by applying the initial detection algorithm, with different values of α, on many intensitygram and magnetogram images. The performance of the algorithm was subjectively analysed for each image. By changing the α value, the number of candidate pixels can be increased or decreased which can affect the outcome of the feature detection process. For intensitygram images an increase in the value of α will decrease the number of sunspot candidates and can cause the missed detection of some sunspots. Also a decrease in this value will increase the number of sunspot candidates and can increase false detections. Changing the value of α affects the detection of active regions in magnetogram images in a similar manner.

After deciding the seeds for active regions a simple region growing algorithm is applied. A 9 ×9 window is placed on every seed and every pixel inside this window that has a similar intensity to the seed's intensity (± 20%) is marked as an active region candidate.
The input and output images in this stage are shown in Figure 3. In Figure 3.c active region candidates with the South polarity are marked with dark pixels and candidates with the North polarity are marked with light pixels.

3.3 Deciding Active Regions and Grouping of Sunspots

After detecting initial candidates for sunspots and active regions, the resulting images are combined to cluster sunspots into groups. Using this method the exact locations of the active regions and sunspots are determined and grouped. This method can be summarized as follows:

1) Get a pixel marked as a candidate ($P_{spotcan}$) on the sunspot candidate image (Figure 4b).

2) If the active region candidate image (Figure 4a) has an active region candidate (P_{actcan}) at the same location, create a new image for active regions and mark it as an active region (P_{act}) with the same pixel value (dark or bright) of P_{actcan} and continue processing, otherwise return to step 1 for processing another $P_{spotcan}$.

3) On the active-region candidate image place a circle on P_{act} with "β" degree radius and mark all the P_{actcan} within this circular region as P_{act} on the newly created active region image.

In this work, the value of β is determined empirically by applying the sunspot-grouping algorithm to five solar image pairs (intensitygrams and magnetograms) that are taken close in time. The value of β is increased gradually from 1 to 15 and it is found using manual inspection that the best grouping performance is achieved when β = 10.

4) After processing all of the $P_{spotcan}$, the created image will show the active regions divided into different polarity regions (Figure 4c). By training and applying a neural network (NN) similar to the one described below, we can decide which polarity regions are coupled with each other and are part of the same active region. Using NN, the different polarity regions that belong to the same active region will be given the same colours and if they are not part of the same group they will be given different colours (Figure 4d).

We used a NN to combine regions of opposite magnetic polarities in order to determine the exact boundaries of sunspot groups. The NN is applied to two opposite polarity regions to decide if they are part of the same active region or not. The NN training vector consists of seven inputs and one output showing the relation between opposite polarity magnetic field pairs.

In order to construct the NN training vector first we calculate the boundaries, area in pixels ($\boldsymbol{A_a}$, $\boldsymbol{A_b}$) and centre of each region in heliographic degrees. We also calculate the distance between the two regions in heliographic degrees ($\boldsymbol{d}$), longitude and latitude difference between the two regions ($\boldsymbol{d_{lon}}$, $\boldsymbol{d_{lat}}$) and the intersecting area between the two regions in pixels ($\boldsymbol{I_{ab}}$). The calculations for input and output members of the training vector are given in Table 1. Figure 5 shows visual descriptions for some of the terms used in this table. Figure 5e is the final image, which is obtained by

ANDing the magnified Figure 5c and Figure 5d (The corresponding area on sunspot candidate image).

The training vector is constructed using nearly one hundred examples. Several experiments are carried to optimise the NN in a manner similar to Qahwaji and Colak (2007). It was found that the best learning performance is obtained with a back-propagation training algorithm and using the following NN topology: Seven input nodes, one hidden layer with eight nodes, and one output node. For more information on NNs please refer to Appendix.

5) Marked regions with the same colour will be counted as a part of same active region and these regions will be combined by filling the gaps between them by marking the spaces with the associated active region colour horizontally and vertically (Figure 4e).

6) Finally this image will be ANDed with the original sunspot candidate image to group the detected sunspots. In this final image every sunspot belonging to the same group will have the same intensity values (Figure 4f).
After deciding the active regions and sunspots, the spots belonging to same groups are marked as detected groups (Figure 6g).All of the stages after pre-processing are shown on Figure 6. The detected groups are then further processed for determining their McIntosh classes.

4. McIntosh Classification of Sunspot Regions

After grouping the detected sunspots, each sunspot group is classified based on the McIntosh classification system which is the standard for the international exchange of solar geophysical data. The classification depends on the size, shape, and spot density of sunspots. It is a modified version of the Zürich classification system, which has improved definitions and added indicators of size, stability, and complexity McIntosh (1990). The general form of the McIntosh classification is *Zpc* where, "*Z*" is the modified Zürich class, "*p*" is the type of penumbra on largest spot, and "*c*" is the degree of compactness in the interior of the group.
In McIntosh (1990) the logical sequence for determining the McIntosh classification and the type of classes for sunspot groups is explained below:

- Computing the modified Zürich class - *Z* :
    - Determine if the group is Unipolar or Bipolar.
    - Determine if a penumbra exist in any of the spots.
    - Determine if the spots with the penumbra are located on one end or both ends.
    - Calculate the length of the group in absolute heliographic degrees.
- Computing the type of penumbra (Largest spot) – *p*
    - Decide if the penumbra of the largest spot is rudimentary or not.
    - Decide if the penumbra of the largest spot is symmetric or not.
    - Calculate the value of the North to South diameter in heliographic degrees.
- Computing the distribution of the sunspot – *c*
    - Determine the compactness of sunspots within the group.
    - Determine if there is a spot with mature penumbra in the group besides the leader and follower.

In this research the same logical sequence is used for determining the McIntosh classification of the sunspot groups.

4.1 Computing the Modified Zürich Class – *Z*

As illustrated earlier, to determine the modified Zürich class we have to find the polarity, penumbra status, and the length of the group.

- The polarity of the sunspot groups is determined based on the separation between sunspots within the group. The largest separation distance between the sunspots within the group is calculated in heliographic coordinates. If there is a single spot or compact cluster of spots in a group and the greatest separation is smaller than 3°, the group is considered to be unipolar; if the separation is higher the group is considered to be bipolar.
- In white-light images, large sunspots have a dark central umbra surrounded by the brighter penumbra. In order to decide if a sunspot has penumbra or not, the mean (μ), standard deviation (σ) of the detected sunspots on the original image is found and a threshold value (T_p) is calculated using Equation (2). Then the detected sunspot pixel values are compared with this threshold value. If the sunspot pixel value is smaller than T_p, it is considered to be part of the umbra; otherwise it is considered to be part of the penumbra.

  $$T_p = \mu - \sigma \qquad (2)$$

  Figure 7c shows the detected umbra and penumbra areas for sunspots. After detecting the umbra and penumbra regions, smaller sunspots within the sunspot group are searched to determine whether they have a penumbra or not.
- The length of the group is calculated by finding the distance separating both ends of the group (*i.e.*, longitudinal extent) in absolute heliographic degrees,

After finding all the necessary information, they are applied to a decision tree to determine the modified Zürich class for the sunspot group.

4.2 Determining the Type of the Largest Spot - *p*

The largest spot in a sunspot group can be classified depending on its type, size, and symmetry of its penumbra (McIntosh, 1990). The penumbra can either be rudimentary (partially surrounds the umbra) or mature (completely surrounds the umbra) and its size is the value of the North to South diameter. A rudimentary penumbra usually denotes a spot that is either forming or decaying. The symmetry of the penumbra depends on the irregularity of the outline associated with this penumbra. A symmetric penumbra is mostly either circular or elliptical in shape. The size of the spot can be easily calculated by finding the difference between its North and South latitudes. However, finding the symmetry and type of penumbra is a real challenge because it depends mostly on the subjective judgment.

In this research we used another NN to determine the symmetry and maturity of each sunspot. The NN training vector consists of nine inputs: kurtosis, standard deviation (σ), mean (μ), skewness, heliographic length (L_{heli}), heliographic diameter (D_{heli}), heliographic area (A_{heli}), penumbra ratio, umbra ratio, and consists of two outputs: symmetry and maturity. Most of these features were used by the authors for the verification of solar features in Qahwaji and Colak (2006a).

The input and output parameters of the training vector are determined as explained in Table 2. For this work we have used the backpropagation neural network because this learning algorithm provides high degrees of robustness and generalisation in classification Kim et al. (2000). To find the optimum NN topology, a large number of learning experiments, in a manner similar to Qahwaji and Colak (2007), were carried out. The performance of the NN is tested after each experiment using the Jack-knife technique. This technique randomly divides the learning data into two sets: a training set containing 80% of the data and a testing set containing the remaining 20%, as explained in Qahwaji and Colak (2007). In this work we have used 100 samples of learning data, each sample consists of nine inputs and two outputs, as explained in Table 2. We found that that the best performance is obtained for the following topology: nine input nodes, one hidden layer with five hidden nodes and two output nodes. For more information on NNs please refer to the Appendix.

After optimisation, the NN is trained. A successful training is achieved if the normalised system error falls below 0.001. After training is completed, the NN is tested with new inputs that were not part of its training examples, in a manner similar to the Jack-Knife technique (Fukunaga, 1990). The output of NN is analysed to determine the maturity and symmetry for each sunspot. If the first output of the NN is higher than 0.5 the sunspot under consideration is assumed to be symmetric otherwise it is considered to be asymmetric. Similarly, if the second output of the NN is higher than 0.5 the sunspot under consideration is assumed to be mature otherwise it is assumed to be rudimentary.

In addition we determined the North to South diameter of the largest sunspots by calculating the longitude and latitude of the upper most and lowermost pixels and then calculating their distances. Depending on the output from the NN the second class of the McIntosh classification system is determined.

4.3 Determining the Sunspot Distribution – *c*

The sunspot distribution depends on the compactness of the sunspot group (McIntosh, 1990). In order to analyze the sunspot distribution within the group, the following steps are followed:

- Find the boundaries of the sunspot group.
- Calculate the area of the group in pixels within the calculated boundaries.
- Calculate the total area of the individual spots in pixels.
- Find the ratio (*R*) of the total spot area to the group area.
- Calculate the number of spots with mature penumbra.

The sunspot distribution type for all the unipolar sunspot groups are “X”. As for the bipolar sunspots, the classification depends on *R*. If *R* is less than 50% then the sunspot group is assumed to be “open” (McIntosh, 1990). If *R* is higher than 50% and the number of spots with mature penumbra is higher than two, the sunspot is assumed to be “compact” otherwise “intermediate”.

5. Implementation and Evaluation

5.1 Practical Implementation of the System

A computer platform using C++ .Net was created for the automated detection and classification of sunspots using SOHO/MDI intensitygram and magnetogram images in the GIF format. A publicly available library: “corona.dll”[1] is used for reading all of the GIF images. The program for training and applying the NN is also created and implemented in C++. The whole system works with 1024 × 1024 images and the detection of sunspots, detection of active regions and classification of sunspot groups

[1] http://corona.sourceforge.net/

takes approximately four seconds per image depending on the complexity of features. The processing time is measured on P4-2.8 GHz PC with 1 GB RAM.

Our system was tested on a total of 103 intensitygram and 103 magnetogram images available from 1 May 2001 until 31 May 2001. Using these images, we created our own catalogue that consists of sunspot groups and their classifications, which will be referred to as the Automated Sunspot Catalogue (ASC) for the rest of this text.

5.2 The Evaluation of the ASC

ASC is compared with the publicly available sunspot catalogues from the National Geophysical Data Center (NGDC)[2]. NGDC keeps records of data from several observatories around the world and holds one of the most comprehensive publicly available databases for solar features. Different observatories provide sunspot classification data at different times. Sometimes there could be three or four "SETs" of data within a single day in a NGDC catalogue, which are provided by several observatories. This makes the NGDC sunspot catalogue suitable for comparison with our ASC. We refer to "SET" as all of the sunspots grouping data that are provided for a specific time in a day. Approximately four SETs are available on ASC per day. Its frequency depends on the availability of MDI images. As ASC is formed by processing 103 images with different dates or times, it has 103 SETs.

For testing the accuracy of ASC, we created a testing program in C++ that will read both catalogues and compare sunspots group data sets according to date, time, location, and classification. This testing program allows us to increase the amount of comparison data, by controlling the time difference for comparing the SETs available in NGDC catalogue and ASC. This program work as follows:

- Read the first SET available from ASC (SET_{ASC}) and calculate its time.
- Calculate the time difference between every SET available on NGDC catalogue (SET_{NGDC}) and SET_{ASC}.
    - If the time difference between SET_{ASC} and SET_{NGDC} is less than the desired time difference (D_T) continue to the next step. Otherwise return to the beginning and do not take this SET_{ASC} into account for comparison.
- Get a sunspot group data from the SET_{ASC} and compare its location with all of the sunspots grouping data on SET_{NGDC}.
    - If any of the sunspot group location SET_{NGDC} and SET_{ASC} matches, mark this group and compare the classifications.
    - If none of the locations match, mark the sunspot group on SET_{ASC} as unmatched.
- Repeat the previous step for the sunspot groups within SET_{ASC}.
- Repeat all of the steps for all of the SET_{ASC} in ASC.

We run our testing program by setting the times of D_T to 30 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, 3 hours, 6 hours, 12 hours, and 1 day. Ideally, the D_T between SETs from ASC and NGDC catalogue should be zero for an accurate comparison but as can be seen from Table 3, even when D_T is made equal to 30 minutes, the number of matching SETs is 19 out of 103. These 19 SETs, corresponding to 179 individual sunspot groups, are not enough for an accurate evaluation. Table 3 shows the results for the evaluation of sunspot grouping. In order to evaluate the grouping performance, the following two error rates are introduced Hong and Jain (1997):

[2] ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/,lastaccess:2007.

- The false acceptance rate (FAR), which is the percentage of a non-sunspot group being detected as a sunspot group.

- The false rejection rate (FRR), which is the percentage of a sunspot group not being detected because it is considered to be a non-sunspot group.

Table 3 shows that the best results for FRR and FAR are achieved when D_T is set equal to 1 hour and 30 minutes. After two hours difference, FRR and FAR rates increases dramatically.

Also, Table 4 shows the evaluation results for our automated McIntosh classification for each D_T setting. In this table, *Z* represents the modified Zurich class, *P* represents the type of largest sunspot, and *C* represents the distribution of the group. The best classification results are achieved up to a maximum of two hours which is logical when we take into account that the change of classification usually takes few hours.

6. Discussions, Conclusions and Future work

6.1 Discussions and Concluding Remarks

To the best of our knowledge, this is the first time that a complete automated system for the detection, grouping, and then classification of sunspots is presented. The system provides reliable and speedy performance. This system processes two types of images simultaneously: SOHO/MDI intensitygram images and magnetogram images. Intensitygram images are processed to detect and gather information about sunspots, while magnetogram images are processed to provide the active region information that is used later to group the detected sunspots.

The system is tested on 103 MDI intensitygram images for the month of May 2001, with a total of 957 sunspot groups and compared with the NGDC sunspot catalogue that are created by solar physicists from different observatories. A program is created using C++ to provide correct evaluation for our system by comparing the sunspots reported in NGDC catalogue with the ones generated in our ASC within the time difference specified. The time difference is increased gradually with the program and the results for comparison are recorded and shown in Tables 3 and 4. These tables show that an accurate evaluation for sunspot grouping can be achieved for a time difference that extends up to one hour 30 minutes and an accurate classification can be achieved for a time difference that extends up to two hours.

If we take into account the one hour and 30 minute time difference for analysing our algorithms, this means that the number of images in our test pool is reduced to 37 with a total of 350 sunspot groups. For the exact period of time, the NGDC sunspots catalogue contains 293 recorded sunspot groups and 261 of them are matched with our 350 detected sunspot groups by simply comparing their locations and timing information. This means that there is nearly a 90% correct match for sunspot groups between the two catalogues and 25% of the groups detected by our algorithms are not reported on the NGDC catalogue. More than 85% of the sunspot groups that are not available on the NGDC catalogue are the sunspot groups with one or two sunspots. We believe that this difference can be caused by:

a) Wrong sunspot detection by our algorithms.
b) Wrong grouping of sunspots by our algorithms.
c) Missed detections of sunspots by solar experts.

Although almost 99% of the detected sunspots are correct, we found that there are some miss detections of very small sunspots (smaller than three pixels). All of the initial sunspot candidates are compared with their corresponding magnetic activity on magnetograms images. This reduces the probability for wrong detection of sunspot candidates. This also shows that most of the errors are caused by wrong grouping of our algorithms and/or miss detections of sunspots by observatories.

Our algorithms clustered some sunspots into separate groups despite the fact that they belong to the same group. This applies in particular to sunspots that are separated by large distances compared to their areas. This causes their magnetic traces to be separated from each other and as a result the NN clusters them as separate groups. Sometimes two or three small sunspots that are part of the same group can be clustered as two or three different sunspot groups.

Also, lack of visibility by ground observatories at the time of sunspots detection and human error (Some small sunspots are very hard to determine by human eye) can cause the miss detections of sunspots. Furthermore, sunspots forming or decaying can be hard to detect. We came across some examples where some sunspot groups are detected by our system in their early development stage and are not reported in the NGDC Sunspot Catalogue until they have matured a little.

An example for wrong grouping and missed detection on NGDC catalogue is shown in Figure 8. In this figure, the detected groups, and classification results on ASC for the images on 01 May 2001 at 00:00 (Figure 8a) and 06:24 (Figure 8b) and also their corresponding SETs on NGDC catalogue is given.

The groups marked as 1a and 1b, which are detected as separate groups on ASC, are actually one group. Our algorithms have not managed to connect these groups and as a result one of the groups is counted as a wrong group on our test results and the one that is closest to group 1 on the NGDC catalogue is counted as the matched group. If we look at the group 6 detected and classified as AXX at 00:00 and as BXO at 06:24 on ASC, we can see that this group is only mentioned as CRO at 07:00 on the NGDC catalogue and there is no information about this group at 00:20 on the NGDC catalogue. This group will be counted as wrong grouping on our test results although it is not.

As can be seen in Table 4, for a maximum of one hour and 30 minutes time difference, out of the 261 matched sunspot groups, the correct classification rates for modified Zurich class (Z), type of largest spot (P), and group distribution (C) are 63%, 47%, and 73% respectively. Also individual matching rates for each of these McIntosh classes are given in Table 5, Table 6 and Table 7. On these tables, we can find the total number of individual classes, their distribution rate among the total number of test groups (261), and also the number and matching rates for individual classes.

Although the modified Zurich class ratio and group distribution ratio results are satisfactory, we can not say the same thing for the type of the largest spot ratio. Deciding the type of the largest spot is a very hard task, even for an experienced solar physicist because it involves subjective judgment on the degree of symmetry and maturity. Our system has a classification rate of 47% for this class but it is hard at this stage to judge whether this is caused by the misclassification of our algorithms, which seems to be more likely, or the misjudgement of observers. In either case, this has to be improved by adding more training examples to the related neural networks or by applying imaging algorithms to detect the geometry of the largest spot (*i.e.*, Hough Transform, *etc.*).

6.2 Future Work

For future work, we plan on improving the grouping and classification rates. Sunspot grouping can be improved by using statistical clustering algorithms for grouping in addition to grouping with the help of the detected active regions from magnetogram images. Classification, especially for determining the type of the largest spot, has to be improved. This can be achieved by a better training of the NN used for deciding the symmetry and type (mature or rudimentary) of penumbra or using other machine learning techniques, such as Support Vector Machines, in a manner similar to Qahwaji and Colak (2007)

We also plan to classify the sunspot groups according to the Mt. Wilson classification which can be done with higher matching ratios when we take into account that the polarity of each sunspot can be easily be determined from the magnetogram images. Our major aim is to combine the output data from this system with a machine learning system, as described in Qahwaji and Colak (2007) to provide an automated platform for the short-term prediction of major solar flares using neural networks and/or support vector machines. More information on this can be found at http://spaceweather.inf.brad.ac.uk/index.html

Acknowledgment

- We would like to thank Dr. Christopher Balch from NWS Space Environment Center, Forecast and Analysis Branch for his advice on sunspot classification, which helped us to improve the quality of this paper.

- Images used in this paper are courtesy of the SOHO/SOI-MDI project of the Stanford Lockheed Institute for Space Research. SOHO is a project of international cooperation between ESA and NASA.

- This work is supported by an EPSRC Grant (GR/T17588/01), which is entitled "Image Processing and Machine Learning Techniques for Short-Term Prediction of Solar Activity".

Appendix:

Neural Networks

The term "neural networks" is used to describe a number of different computational models intended to imitate biological neurons. These models consist of artificial neurons (Figure 9) connected to each other, where the connections, also known as synaptic weights, are used to store the knowledge.

A neural network can consists of numerous artificial neurons that are arranged into layers. Each layer is interconnected with the layer before and after it (Figure 10). The input layer is the first layer and it receives external inputs, while the outputs are provided by the last layer, which is also called the output layer. The other layers between the input and output layers are called hidden layers. There are two basic NNs topologies: Feed-forward and feed-backward. In the feed-forward model information are fed from the input layer toward the output layer and the output of each layer is used as the input to next layer. In feed-backward model the output from a layer can be used as an input to itself or to previous layers.

NNs can be trained using supervised and unsupervised learning algorithms. In unsupervised learning, the network is provided with the inputs only and the system decides how to cluster the input data. The training of the network using inputs and their corresponding outputs is called supervised learning. In supervised learning, each sample in the training set specifies all inputs, as well as their desired outputs. A set of examples used for training is called "training set" and samples from the training set are chosen and presented to the network one at a time. For each sample, the outputs generated by the network and the desired outputs are compared. After processing all of the samples in the training set, the neural weights are updated to reduce the error.

Reference

Curto, J.J., Blanca, M., Solé, J.G. 2003, *Solar Image Recognition Workshop* ,Brussels.

Fukunaga, K. 1990, *Introduction to Statistical Pattern Recognition,* Academic Press, New York,220.

Hathaway, D., Wilson, R.M., Reichmann, E.J. 1994, *Solar Phys.* **151**, 177.

Hong, L., Jain, A. 1997, IEEE Trans. Pattern Analysis Machine Intelli., **20**, 1295.

Künzel, H. 1960, Astronomische Nachrichten 285, 271.

Kim, J., Owat, A., Poole P., Kasabov N. 2000, *Chemometrics Intelligent Laboratory Systems*, **51**, 201.

McIntosh, P.S. 1990, *Sol. Phys.*, **125**, 251.

Meeus, J. 1998, *Astronomical Algorithms - Second Edition*, Willmann-Bell, Inc., Virginia.

Nguyen, S.H., Nguyen, T.T., Nguyen, H.S. 2005, *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Slezak, D., Yao, J., Peters, J.F., Ziarko, W., Hu, X. (eds.), Springer Berlin, Heidelberg. **3642**, 263.

Phillips, K.J.H. 1992, *Guide to the Sun*, Cambridge University Press, Cambridge.

Qahwaji, R., Colak, T. 2006a, International Journal of Imaging Systems & Technology **15**, 199.

Qahwaji, R., Colak, T. 2006b, The International Journal of Computers and Their Applications **13**, 9.

Qahwaji, R., Colak, T. 2007, *Solar Phys.*, **241**, 195.

Sakurai, K. 1970, *Planet Space Sci.,* **18**, 33.

Scherrer, P.H., Bogart, R.S., Bush, R.I., Hoeksema, J.T., Kosovichev, A.G., Schou, J., Rosenberg, W., Springer, L., Tarbell, T.D., Title, A., Wolfson, C.J., Zayer, I., Akin, D., Carvalho, B., Chevalier, R., Duncan, D., Edwards, C., Katz, N., Levay, M., Lindgren, R., Mathur, D., Morrison, S., Pope, T., Rehse, R., Torgerson, D. 1995, *Solar Phys.,* **162**, 129.

Severny, A.B. 1965, *Stellar and Solar Magnetic Fields, International Astronomical Union. Symposium no. 22*. Lust, R. (ed.), North-Holland Pub. Co., Amsterdam, 358.

Warwick, C.S. 1966, *Astrophys. J.,* **145**, 215.

Zharkov, S., Zharkova, V., Ipson, S., Benkhalil, A. 2004, *Knowledge-Based Intelligent Information and Engineering Systems, Pt 3, Proceedings*, **3215**, 446.

Figure 1: Images showing Stage-2 processing. “a” is the cleaned (Stage-1) magnetogram image in Heliocentric-Cartesian coordinates, “b” is the image in Carrington-Heliographic coordinates, “c” is the image mapped back to Heliocentric-Cartesian coordinates with new values, “d” is the difference between “a” and “c”.

Figure 2: Images showing results of Stage-2 processing. “a” and “b” are magnetogram images with 25 hours difference. “c” is the resulting image when Stage-1 and Stage-2 processing is applied to “a”. “d” is the resulting image when Stage-1 and Stage-2 processing is applied to “b” using the solar coordinate and radius values from “a”. White line going through images provided for showing the rotation on images.

Figure 3: The results of initial detections for sunspots and active regions. “a” and “b” are magnetogram and intensitygram images respectively. “c” is the image showing active region candidates. “d” is the image showing sunspot candidates.

Figure 4: Deciding active regions and sunspot groups. “a” and ”b” are selected areas from active region and sunspot candidate images, “c” is the resulting image after growing sunspot candidates inside active region candidates, “d” is coloured regions after applying NN, “e” is the final active regions, and “f” is the image showing final sunspots with the ones belonging to same group have same intensity values .

Figure 5: Visual descriptions of terms used for constructing NN training vector. “a” is the resulting image after growing sunspot candidates inside active region candidates, “b” is showing some terms used, “c” is one of the magnified regions from “a” and ”d” is the corresponding area on sunspot candidate image, “e” is the result of ANDing “c” and ”d”; sunspots intersecting two opposite polarity regions are shown on this image.

Figure 6: Stages in detecting and grouping sunspots. “a” and “b” are magnetogram and intensitygram images respectively. “c” is the image showing active region candidates. “d” is the image showing sunspot candidates. “e” is the resulting image when “d” and “c” combined using region growing. “f” is the image created by applying NN to regions on image “e” and “g” is the final image that shows sunspot groups detected.

Figure 7: Deciding penumbra and umbra of spots on detected sunspot regions. “a” is the image showing detected sunspot groups, “b” is the magnified area from original image and ”c” is the same area showing penumbra and umbra areas detected by our algorithms.

Figure 8: Comparison of grouping and classification results on ASC and NGDC catalogue.

Figure 9: An artificial neuron, where “i” represents inputs, “w” represents weights.

Figure 10: A multi-layered neural network.

Table 1: Inputs and output for NN training vector for active region decision.

INPUTS	DESCRIPTION
Min(A_a , A_b) / Max(A_a , A_b) I_{ab} / A_a I_{ab} / A_b d_{lon} / d d_{lat} / d d / 180 0.1 or 0.9	Ratio of the smallest area to biggest area of regions Ratio of intersecting area to area of first region Ratio of intersecting area to area of second region Ratio of the longitude difference between regions to distance between regions Ratio of the latitude difference between regions to distance between regions Ratio of distance between regions to 180 degrees. If two regions are intersected by same sunspot candidate it is 0.9 otherwise 0.1.
OUTPUT	DESCRIPTION
0.1 or 0.9	If two regions are part of the same active regions it is 0.9 otherwise 0.1

Table 2: The input and output parameters involved in the NN training to determine the sunspot penumbra type.

INPUTS	DESCRIPTION
Kurtosis σ /255 μ /255 *Skewness* L_{heli} D_{heli} A_{heli}	The distribution measurement that shows the peakedness (broad or narrow). Normalized standard deviation value of the sunspot. Normalized mean value of the sunspot. The distribution measurement that shows distortion in a positive or negative direction. The heliographic length of the sunspot. N-S diameter of the sunspot in heliographic degrees. Area of sunspots in heliographic degrees.
P_{pen} / A_{pixel}	Ratio of number of pixels that are part of the penumbra to total number of pixels on sunspot.
P_{umb} / A_{pixel}	Ratio of number of pixels that are part of the umbra to total number of pixels on sunspot.
OUTPUT	DESCRIPTION
0.1 or 0.9	If the sunspot is symmetric output is 0.9, if it is asymmetric output is 0.1.
0.1 or 0.9	If the sunspot is mature output is 0.9, if it is rudimentary output is 0.1.

Table 3: Evaluation of Sunspot Grouping for Different D_T Values.

D_T (hour)	Total number of SETs	Number of Matched SETs	Total Sunspot Groups on NGDC SETs	Total Sunspot Groups on ASC SETs	Total Number of Matched Sunspot Groups	FRR	FAR
0.5	103	19	155	179	138	10.9%	22.9%
1	103	25	195	225	174	10.8%	22.7%
1.5	103	37	293	350	261	10.9%	25.4%
2	103	47	389	439	330	15.2%	24.8%
3	103	57	529	535	413	21.9%	22.8%
6	103	70	738	656	505	31.6%	23.0%
12	103	96	847	898	605	28.6%	32.6%
24	103	103	948	957	618	34.8%	35.4%

Table 4: Evaluation of McIntosh Subclasses for Different D_T Values.

D_T (hour)	Total Number of Matched Sunspot Groups	Number of Matched SETs	Correct Z Ratio	Correct P Ratio	Correct C Ratio
0.5	138	19	64.5%	47.1%	72.5%
1	174	25	63.8%	43.7%	73.0%
1.5	261	37	62.8%	47.1%	73.2%
2	330	47	63.3%	47.0%	75.2%
3	413	57	62.0%	45.5%	75.8%
6	505	70	58.8%	44.6%	73.3%
12	605	96	56.2%	42.6%	69.6%
24	618	103	53.9%	42.4%	67.6%

Table 5: Evaluation of modified Zürich Class classification for one hour 30 minutes time difference.

Modified Zurich Class	A	B	C	D	E	F	H
Matched	14	10	21	25	18	13	63
Total	19	24	38	51	36	17	76
Distribution Rate	7.3%	9.2%	14.6%	19.5%	13.8%	6.5%	29.1%
Matching Rate	73.7%	41.7%	55.3%	49.0%	50.0%	76.5%	82.9%

Table 6: Evaluation of largest spot class classification for one hour 30 minutes time difference.

Largest Spot Class	X	R	S	A	H	K
Matched	37	1	44	32	0	9
Total	43	10	109	85	2	12
Distribution Rate	16.5%	3.8%	41.8%	32.6%	0.8%	4.6%
Matching Rate	86.1%	10.0%	40.4%	37.7%	0.0%	75.0%

Table 7: Evaluation of sunspot distribution class classification for one hour 30 minutes time difference.

Sunspot Distribution Class	X	O	I	C
Matched	85	100	5	1
Total	95	132	32	2
Distribution Rate	36.4%	50.6%	12.3%	0.8%
Matching Rate	89.5%	75.8%	15.6%	50.0%