For SIP Topical Issue.

Automated Boundary-Extraction and Region-Growing Techniques Applied to Solar Magnetograms

R. T. James McAteer
National Research Council Research Associate, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, U.S.A. (j.mcateer@grasshopper.gsfc.nasa.gov)

Peter T. Gallagher
Department of Experimental Physics, University College Dublin, Dublin 4, Ireland.

Jack Ireland and C. Alex Young
L-3 Communications GSI, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, U.S.A.

Abstract. We present an automated approach to active region extraction from full disc MDI longitudinal magnetograms. This uses a region- growing technique in conjunction with boundary-extraction to define a number of enclosed contours as belonging to separate regions of magnetic significance on the solar disc. This provides an objective definition of active regions and areas of plage on the Sun. A number of parameters relating to the flare potential of each region are discussed.

1. Introduction

Solar physicists are currently confronted by a daunting challenge which many other fields would be glad to have; that of a overwhelming data load. Within a few years the data volume from solar observatories will have increased by over three orders of magnitude. For example, compare the data volume from the Solar and Heliospheric Observatory (SoHO; Domingo, Fleck, and Poland, 1995) of 1Gbyte per day to the 1Tbyte per day projected data volume of the Solar Dynamics Observatory (SDO). Improved computational resources (both software and hardware) will be vital to maximise the science output from these new data. Simultaneously, the need to study this data in near real-time for solar activity predictions is becoming paramount to many commercial (e.g., airlines), telecommunication (e.g., satellites) and governmental (e.g., astronaut safety) activities. As such, a system to automatically identify, classify, and assess the flaring potential of active regions quickly and reliably is highly propitious.

Several centers worldwide already provide a service of this type. The Space Environmental Centre (SEC, part of NOAA; http://sec.noaa.gov/) provides regular solar activity updates, especially when a large (>M5) flare occurs. The Max Millennium program of solar flare research (http://solar.physics.montana.edu/max_millennium) in conjunction with the Active Region Monitor (http://www.solarmonitor.org; Gallagher *et al.*, 2002) and the Solarsoft

Latest Events page (http://www.lmsal.com/solarsoft/last_events/) provide daily flare forecasts. However, these provide data, and flare-related parameters, based on a mostly subjective system. There have been attempts at automating some aspects of this process. The European Grid for Solar Observations (http://www.egso.org/) provides an archive of automatically identified active regions based on Hα and Ca II K data (Zharkova *et al.*, 2005). Worden, White, and Woods (1998) also discuss the automated detection of features in Ca II K data while Verwichte (2003) focuses on active region extraction from extreme-ultraviolet images. Other work uses various image processing techniques: Pap *et al.* (1997) discuss a Bayesian approach to identify solar features; Bratsolis and Sigelle (1998) discuss a mean field fast annealing algorithm to extract active regions from complex Hα images; automated solar flare tracking routines using image-processing techniques are in place at Big Bear Solar Observatory (Qu *et al.*, 2004).

There are two main issues to address in designing any automated classification of a physical system. The first of these is the *dataset*, which must possess the following qualities.

– Data Resolution & Coverage - Physical system is assumed not to change significantly below the spatial and temporal resolution of the data (the physical system is completely sampled). All data is free of major distortions and the dataset covers any know spatial and temporal dependence.

– Spectral Window - This relates to the need to cover the entire physical system (i.e, over the entire energy range). Initially the spectral window which provides the best classification by eye should be used.

– Data Availability - Data must be easily, readily available and of a consistent quality.

When studying active region classification, these factors naturally lead to the choice of the Michelson Doppler Imager (MDI; Scherrer *et al.*, 1995) as the main data provider. This routinely provides full disc photospheric longitudinal magnetograms at a cadence of 96 min, at a resolution of 3.92". Ground-based data may be of better spatial resolution but require significant image reconstruction and will not be available at all times. Current vector magnetograms provide better quality data (including the entire magnetic field vector), but only over smaller areas of the Sun and often require CPU-intensive image reconstruction. MDI 96-min data is readily available over a large portion of the solar cycle (nearly 9 years of data to date) and is full disc, hence we do not introduce artificial biases caused by analysing data from a position on the disc or from a particular period in the solar cycle.

The second main issue to address is that of pure image-processing; which *attributes* best describe the physical system and how may they be parameterised? An important issue here is that high-order statistics (Lawrence *et* *al.*, 1996; Abramenko *et al.*, 2003) are typically CPU intensive but simple parameters may not describe the physical system completely. Hence the need for near real-time flaring assessment must be balanced against the need to adequately describe the system. Vlahos and Georgoulis (2004) and Lu (1995) have shown that the solar photospheric magnetic field may be in a state of non-linear self-organised criticality, so high-order statistics are necessary. Hence an incomplete description of the physical system will make the the classification less than satisfactory.

This paper provides an initial attempt at supplementing the works described above by describing an automated active region selective extraction routine. The aim is to use this routine on a learning dataset to classify regions of magnetic significance (mainly active regions, but also plage) and ultimately provide flare potential of active regions in near real-time. In this paper we concentrate on the main image-processing parts of the system and extracting low-order statistics only, leaving higher-order statistics (McAteer *et al.*, 2005) and classification (Nyugen *et al.*, 2004) to future work. In § 2 we discuss the three main parts of the routine; instrumental effect corrections, boundary extraction, and region growing. These are applied to an example magnetogram in § 3 and resulting statistics are discussed. The regions defined as magnetically significant, and statistics arising from these, are compared to NOAA active regions. Finally we reflect on the many applications and extensions of this work in § 4.

2. Active Region Selective Extraction

The main thrust of this algorithm is to automatically and quantitatively identify active regions on the solar disc, while using as few arbitrary parameters as possible. There will always be a need to introduce some parameters of a partially arbitrary nature (due to the finite resolution of the data), but any of these parameters should be both as well-considered (i.e., studied over large range of values) and as free (i.e., large active regions may require different parameter values compared to small plage regions) as possible. It is also important to only carry out the minimum necessary corrections to the data to limit artifacts.

Several times a day the MDI instrument measures the full disc longitudinal magnetic field using a series of polarizers and scanning a 9.4 nm bandpass across the Ni 676.8 nm absorption line. The inversion of this scan is carried out by the MDI team, and the resulting calibrated 5-min averaged magnetograms are available from the MDI data centre (http://soi.stanford.edu/). The example shown throughout this paper are data from a day selected at random, July 11 2002 (Figure 1).

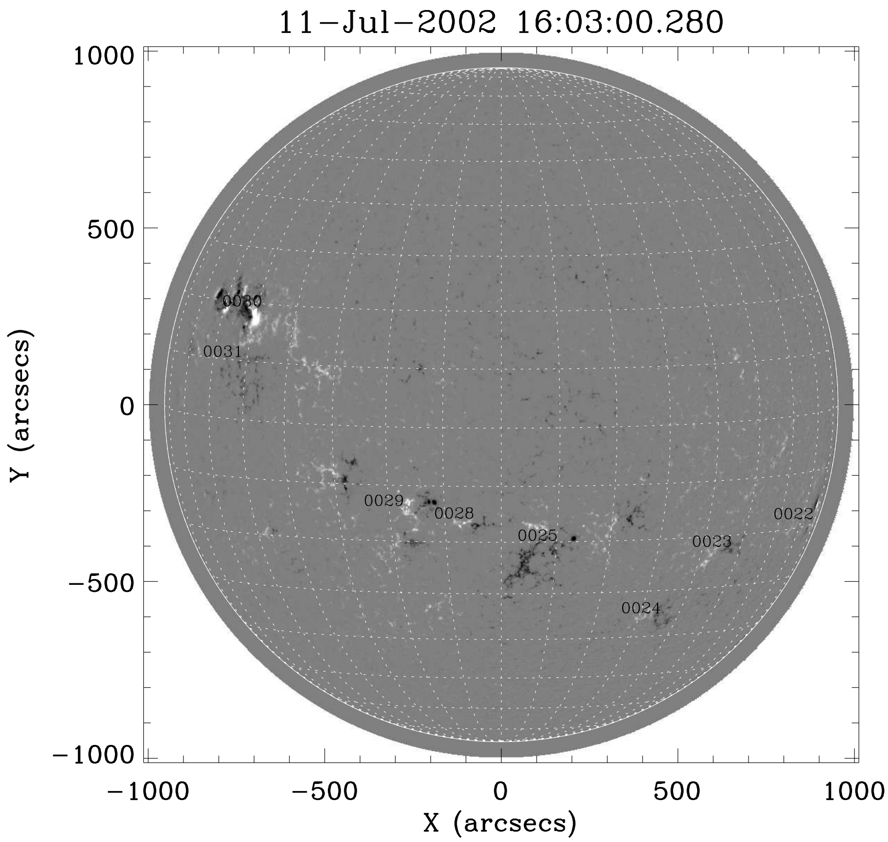

Figure 1. Example full-disc MDI longitudinal magnetogram, overlaid with heliographic lines at 10^o separation and NOAA active region numbers.

2.1. Instrumental Effects

As with any magnetogram the field strength assigned to each MDI pixel is actually the flux averaged over the resolution element. Hence different instruments will measure different field strengths from the same point on the solar disc. In an effort to combat this Berger and Lites (2003) compare MDI full disc magnetograms to data from the Advanced Stokes Polarimeter and discovered a systematic underestimation by MDI. Importantly, this underestimation becomes non-linear at high field strengths, hence will have a large effect on any statistics calculated from the data. Following the suggestion of Berger and Lites (2003) (and after Green *et al.*, 2003) we apply a multiplicative factor of 1.45 for all pixels with an absolute field strength less than 1200G and a multiplicative factor of 1.9 for all pixels with an absolute field strength

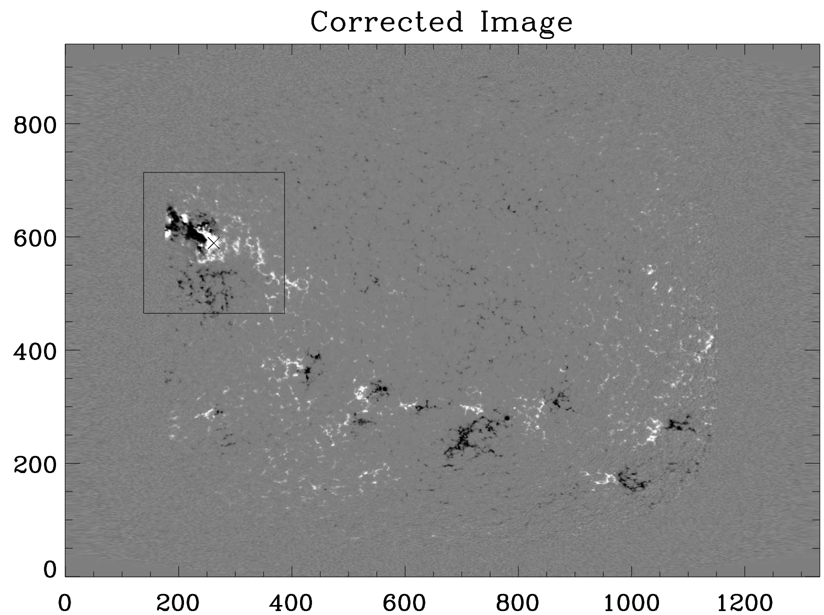

Figure 2. The result of correction for instrumentation effects. The equal-area cylindrical projection ensure that all pixels close to the centre of this data have equal size ($\sim$2 Mm^2) but this will be significantly large near the poles. Pixels close to the limb were replaced by MDI noise prior to the projection correction. The pixel of largest absolute field is marked with an 'X' and becomes the seed for the first region. A sub-window of size 250$\times$250 pixels centred on this seed pixel is extracted.

greater than 1200G (This factor of 1.9 is an upper bound, which accounts for MDI saturation).

Assuming a purely radial field vector we then apply the following cosine correction factor to each pixel, i,

$$B_{corrected}(i) = \frac{B_{MDI}(i)}{\sin(\arccos(\frac{r(i)}{r_{sun}}))}, \tag{1}$$

where r(i) is the heliocentric angle at each pixel, and r_{sun} is the heliocentric radius of the solar disc. At this stage, all pixels at a heliographic angle greater than 60^o were replaced with MDI noise. This purely because these pixels will create false high-valued streaks in the projection-corrected data which collapses the routine.

In order to correct for the projection effects which invariably exist in any solar data, we apply an cylindrical equal-area mapping as created by Lambert in the late eighteenth century. This was chosen so as to minimise the distortion out at latitudes less than 40^o (within which most active regions occur). Under

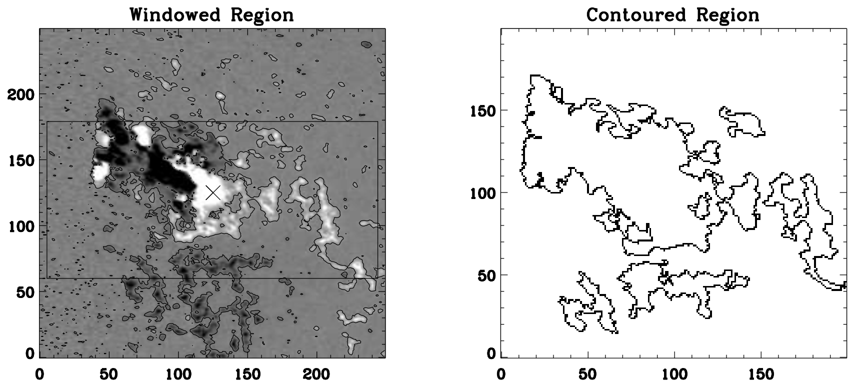

Figure 3. Boundary Extraction. Left: The windowed section of the corrected data, contoured at $\pm$ 50G. Both active region and quiet Sun pixels are selected. Right: Closed contours which exist at least partially inside the rectangle (of size X=2$\times A_{thresh}$ and Y=A_{thresh}) centred on the seed pixel 'X', and consisting of at least C_{thresh} pixels define the active region boundary.

this projection all pixels are remapped via

$$X_i = R \times x_i, \tag{2}$$

$$Y_i = R \times \sin(y_i), \tag{3}$$

where (x, y) is the longitude and latitude at each pixel, i, in the full disc data, R is the radius of the solar disc (in pixels), and (X, Y) is the new co-ordinate system (axes of Figure 2). Essentially this projects the solar disc onto a cylinder, and unwraps this cylinder to provide data where lines of equal longitude are parallel and equi-spaced and lines of equal latitude are parallel, with smaller separation at the poles. Hence each pixel near the equator in the corrected image corresponds to an area of size $\sim$2 Mm^2 on the solar disc. The reader is directed to Bugayevskiy and Snyder (1995) and Snyder (1993) for further details.

As the corrected image is substantially larger than the input full disc image, missing data is interpolated using a cubic spline. DeForest (2004) has shown that this simple interpolation produces severe aliasing at high spatial frequencies. However as we are dealing with large scale structures, and bearing in mind the need to keep the routine as fast as possible, we stick to interpolation. At this stage flare event lists and NOAA active region positions, provided by the SEC, are also read in and all flare and active region positions are converted to this new projected co-ordinate system.

2.2. BOUNDARY-EXTRACTION

The most difficult part of extracting any subset of pixels from an array is in deciding the rules which define membership to this subset. In this work, membership is initially defined by a boundary extraction algorithm. The first step in the routine is to select the pixel of largest magnetic field strength in the corrected image. This is designated as the seed pixel for the first active region. A window of 250×250 pixels is extracted centred on this seed pixel.

It is well documented that any feature extraction routine based solely on simple thresholding is highly sensitive to the threshold chosen (Gonzalez and Woods, 2002). However an initial threshold must be chosen, if only to remove pixels severely affected by noise; we choose an absolute threshold of B_{thresh}. Rather than simply masking these pixels, contours are drawn on the image at this chosen field strength. Figure 3 shows that this creates contours around both active region and quiet-Sun areas. The simplest method of removing the quiet-Sun areas is to increase the threshold, but this also removes active region pixels. Instead, we introduce two free morphological parameters to preserve as much active region at as low B_{thresh} as possible, whilst removing the quiet-Sun areas. Any closed contour consisting of more than C_{thresh} pixels, and partially existing within a rectangle of size scale A_{thresh} is defined belonging to the same active region as the seed pixel. The size scale is defined twice as large in longitude as latitude to account for the fact that active regions tend to be larger in longitude and bunched in latitude. Hence, the main thrust here is to include both polarities (and any associated plage) of an active region within the defined boundary whilst separating any regions which may be close by. The combination of values finally chosen were B_{thresh}=50G, A_{thresh}=C_{thresh}=20–50. We note that B_{thresh} is well above the expected noise level of 9G (Domingo, Fleck, and Poland, 1995) and similar to the value chosen by other authors using MDI data (Muenier, 1999). A_{thresh} and C_{thresh} were allowed to vary according to active region size; for a region consisting of n closed contours each of length, l_n, this was achieved by taking one fifth of the average of the ten longest contours,

$$A_{thresh} = C_{thresh} = \frac{\Sigma_{i=1}^{10} l_i}{50}. \tag{4}$$

It is important to realise that these values were not arbitrarily forced upon the data. Instead a large range of values were searched and these values selected occurred at a plateau in A_{thresh}–B_{thresh}–C_{thresh} phase space where small changes in any parameter had a minimum effect on the output (McAteer *et al.*, 2005). Hence, in effect, the morphology of the data itself was used to extract the optimum values of A_{thresh}, B_{thresh} and C_{thresh}.

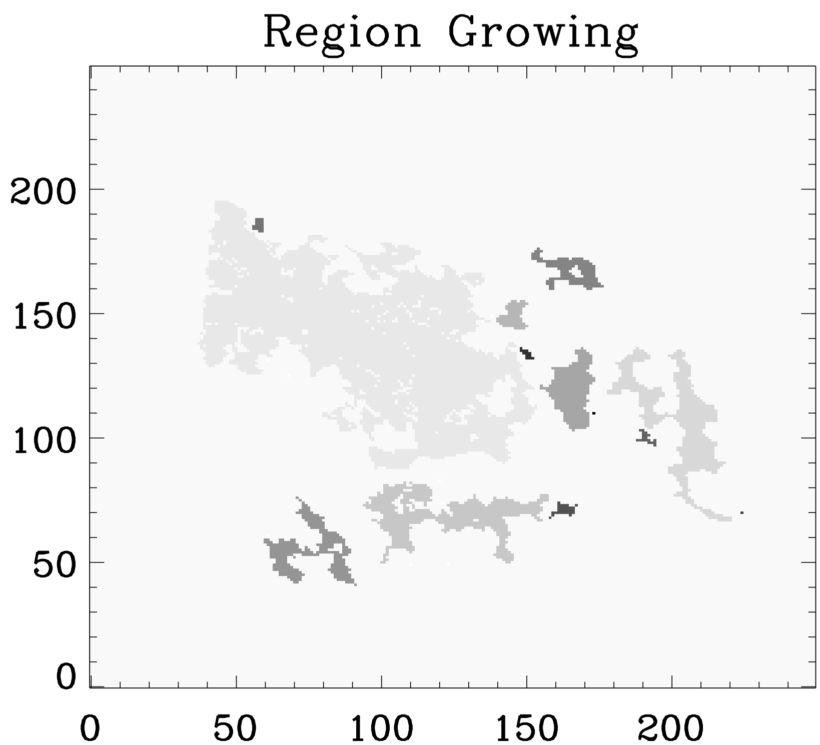

Figure 4. A fully grown region. Each closed contour is displayed as a different intensity, with the largest contour displayed at lowest greyscale. All non-white pixels constitute the fully grown active region.

2.3. REGION-GROWING

At this stage the two main components – the seed pixel and the boundary – of any region-growing routine are already defined. All 8 pixels contiguous to the seed are considered as candidates for the region-growing. Each candidate is accepted as belonging to the same active region as the seed pixel if it obeys the following three conditions.

- It has a field strength above the chosen threshold of 50G
- It exists inside a closed contour
- It does not already belong to an predefined active region

The new active region pixels then become seed pixels and all pixels contiguous to these, and not already defined as definitely belonging, or not, to the region, become the new candidate set. The routine is repeated for each new candidate set until all pixels inside closed contours have been defined as either belonging, or not, to the region (Figure 4).

2.4. Region Selection

The completed region is removed from the corrected image and the routine is repeated using the pixel with largest remaining field strength as the new seed pixel. When each region is fully defined, it is removed and the routine repeated again, until no regions remain. This occurs when either no remaining pixels have a field strength above the threshold of 50G or no closed contour of sufficient perimeter length ($> C_{thresh}$ pixels) remains.

3. Results & Discussion

Figure 5 displays the final output of the algorithm for July 11 2002. The top half shows the 13 selected regions of magnetic significance (numbered in order of decreasing maximum field strength) over-plotted on the projection-corrected image. The bottom half shows each region in a different greyscale value and the position of the NOAA numbered active regions. The routine has successfully identified and separated each region, however two effects are immediately striking and will contribute to the false rejection and false acceptance rate.

First of all the routine has not selected any limb regions. This is because the correction for projection effects fails close to the limb, resulting in wild data values, and the routine will collapse. Hence all limb pixels (at greater than 60^o heliographic) are replaces with MDI noise at the initial stages. There is no real solution to this problem, although we may include a *memory* component in future version. This will allow a region to be followed closer to the west limb, on the basis that the routine will be able to predict where a region should be, based on its previous position. It may also be able to predict when a region will return on the east limb.

Secondly, the routine has identified some regions of magnetic significance which are not NOAA numbered active regions. This effect can be reduced by introducing another arbitrary parameter (e.g., all active regions must contain some field strength above 500G, or all active regions must consist of at least 50 pixels). We prefer not to introduce another arbitrary parameter, rather to over-select magnetic regions and then classify these. These ‘false acceptance’ regions are either (i) plage, (ii) decaying regions or (iii) emerging regions. Of these three possibilities it is trivial to recognize plage (see § 3.1). A memory component will separate the decaying regions from the emerging regions. Any newly emerged and rapidly growing regions will be of most interest in judging solar flare potential.

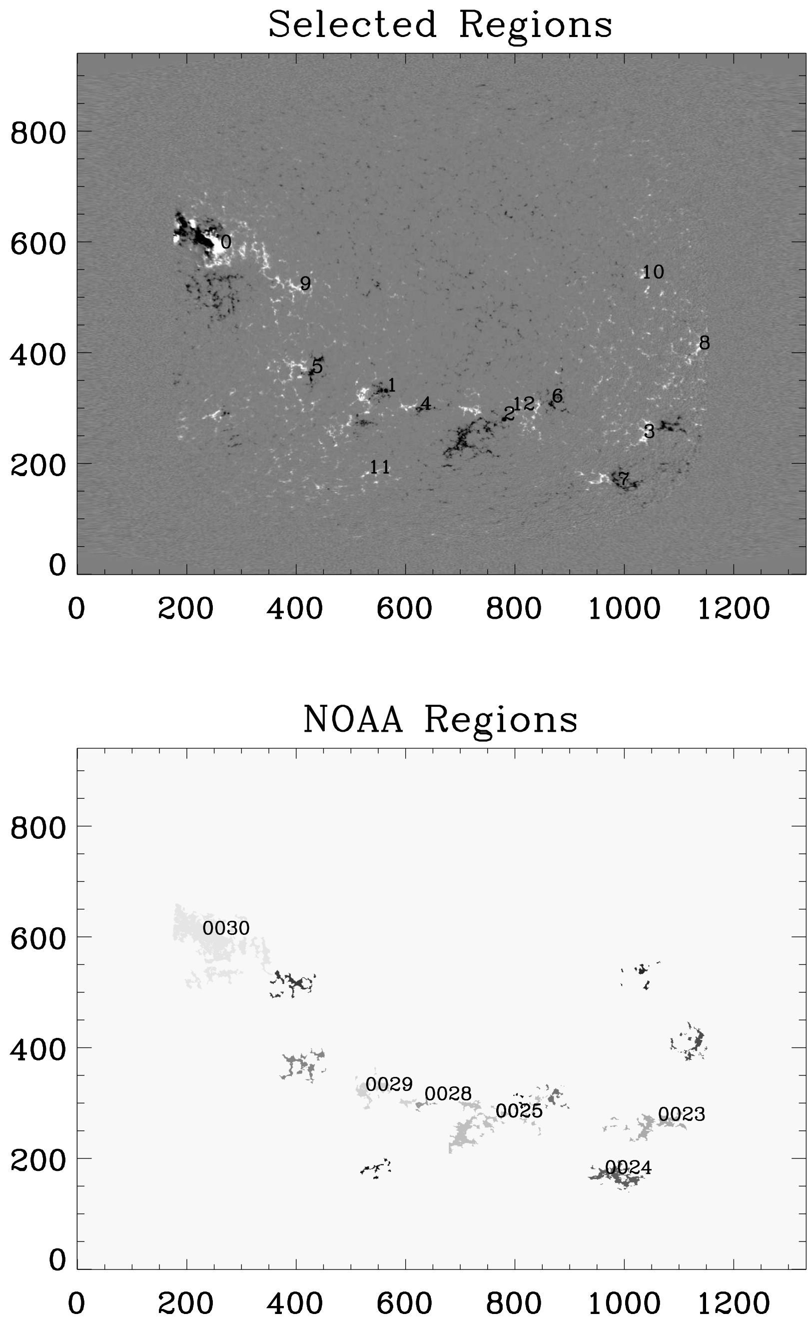

Figure 5. Final selected regions. Top: The selected regions(0–12) are marked on the projected data of Figure 2. Bottom: Each region is displayed at a different greyscale along with its identified NOAA number

Table I. Statistics from the final selected regions. These are position (in pixel coordinates on the corrected image), size (in Mm^2) of the positive and negative parts, total size (in Mm^2), flux (with filling factor of unity) of the positive and negative parts, total flux, total absolute flux, number of contours of mixed polarity, and total number of contours.

Num	Position	Size (Mm^2)			Flux ($\times 10^{15}$Mx)				Contours	
	(X,Y)	(+)	(-)	Total	(+)	(-)	total	total abs	Mixed	Total
0	262,589	7411	8685	16096	100582	-119236	-18654	219818	1	20
1	565,331	1578	806	2383	3269	-1017	2252	4286	1	15
2	780,279	1486	3689	5174	2485	-18855	-16370	21340	2	24
3	1036,247	1684	1262	2946	3197	-1954	1243	5151	1	10
4	628,298	0	392	394	0	-202	-202	202	0	4
5	429,364	1190	870	2060	1616	-1038	578	2654	1	9
6	868,311	20	741	760	0	-627	-626	627	1	10
7	989,161	739	1893	2632	624	-3704	-3080	4327	3	19
8	1137,407	1368	12	1380	1955	0	1955	1955	2	25
9	407,514	1595	0	1597	2887	0	2887	2887	0	7
10	1031,535	498	45	543	284	-1	283	286	1	7
11	532,183	512	0	514	258	0	258	258	0	5
12	794,297	147	0	149	19	0	19	19	0	5

3.1. Low-order statistics

Once the regions are identified and selected, any number of image-processing tools can be applied to describe the system. Here, we only produce a number of low-order statistics, given in Table I. The positions (X,Y) are in pixel coordinates of the corrected image and are used to identify these regions with their NOAA counterparts. The calculated areas (positive field, negative field, total active region) are in physical values (Mm^2), compared to NOAA values which are in millionths of the solar disc. The flux values are calculated using a filling factor of one (i.e., this value is maximum flux. As this flux is calculated from the longitudinal photospheric field, the actual values calculated should be treated with extreme caution; the relative comparison of flux between regions (and between days) will be more useful. The comparison of positive and negative field also allows us to identify plage regions (unipolar) and regions of large flux imbalance (Green *et al.*, 2003).

The final two columns in the table describe how many of the closed contours defining the region contain mixed polarity and, as such, has two possible uses. Mixed regions (when both polarities are in close proximity) are much more likely to flare than simple bipolar regions. Secondly, we can use this in conjunction with MDI broadband white light images to automatically assign delta components in the Mount Wilson classification system. Some care must be taken here, as some regions of mixing may be due to a limb effect, where the inversion of the spectra may make positive field appear negative and vice versa (see regions 6,8,10). This may be addressed by examining the gradient across these fake neutral lines, which should be much smaller than across real neutral lines.

From Table I it is immediately obvious that the largest (columns 3–5), most complex (columns 6–11) region is region 0. Indeed a comparison with the flare list shows this is NOAA 10030 which produced 2 M-class and 5 C-class flares within 24 hours of this image. The second largest region, 2, corresponds to NOAA 10025 which produced 2 C-class flares. The other NOAA regions (10023–3, 10024–7, 10028–4, 10029–1 were simple non-flaring beta regions. The remaining identified regions are weak unipolar plage regions (6,8–12) and one unnumbered beta region (5).

4. Future Work

We have demonstrated a fully automated routine to extract regions of magnetic significance from full-disc magnetograms. The routine successfully identified most NOAA numbered regions with a few exceptions. We envisaged this as pure image-processing task with no *a priori* physical knowledge. However, both the false acceptance and false rejection rate can be lowered by introducing a memory component into the routine. We also intend to use a similar automated extraction routine on MDI broadband white light images in an attempt to automatically assign Mount Wilson and/or McIntosh classes to each region. Similar routines may also be extended to data from other instruments(e.g., Extreme-ultraviolet Imaging Telescope; Verwichte, 2003) to provide a 3-dimension topology of each region, or to classify other solar features (e.g., filaments).

As our aim is to provide near real-time data analysis, computation time is an important factor in our routines and limit our choice of data and techniques. However, our routines are generic and modular and hence once better data (e.g., MDI high resolution images, vector magnetograph data) and routines (e.g, multiscale) are available and process rapidly, we will be including them in our analysis. The long term goal is to use these routine, and others, on a large dataset (e.g., every MDI full disc magnetogram) to build up a new classification system based on objective methods, which will supplement the subjective systems currently in place. All data, resulting statistics, and routines will be made available online at the Active Region Monitor.

Acknowledgements

JMA acknowledges an NRC Research Associateship award for this work which was carried out in the RHESSI group at GSFC. Useful discussions and programs regarding projection effects were obtained from Tom Berger, Satoshi Morita and Craig DeForest. The flare data was provided by the Space Environmental Centre. Data was gathered from the MDI data centre at Stanford. All authors are grateful to the participants of the 2nd Solar Image Processing Workshop for a productive workshop.

References

Abramenkeo, V.I., Yurchyshyn, V.B., Wang, H., Spirock, T.J., and Goode, P.R.: 2003, *Astrophys. J.*, **597**, 1135

Berger, T., and Lites, B.W.: 2003, *Solar Phys.*, **213**, 213

Bratsolis, E., and Sigelle, M.: 1998, *Astron. Astrophys. Suppl. Ser.*, **131**, 371

Bugayevskiy, L.M., and Snyder, J.P.: 1995, *Map Projections A Reference Manual*, Taylor & Francis, Inc., Philadelphia

DeForest, C.E.: 2004, *Solar Phys.*, **219**, 3

Domingo, V., Fleck, B., and Poland, A.I.: 1995, *Solar Phys.*, **162**, 1

Gallagher, P.T., Moon, Y.-J., and Wang, H.: 2002, *Solar Phys.*, **209**, 171

Gonzales, R.C. and Woods, R.E.: 2002, *Digital Image Processing*, 2nd ed. Prentice-Hall Inc., New Jersey

Green, L.M., Demoulin, P., Mandrini, C.H., van Driel-Gesztelyi, L.: 2003, *Solar Phys.*, **215**, 307

Lawrence, J.K., Ruzmaikin, A.A., and Cadavid, A.C.: 1996, *Astrophys. J.*, **465**, 425

Lu, E.T.: 1995, *Astrophys. J.*, **446**, L109

McAteer, R.T.J., Gallagher, P.T., and Ireland, J.: 2005, *Astrophys. J.*, submitted

Muenier, N.: 1999,*Astrophys. J.*, **515**, 801

Nyugen, T.T., Willis, C.P, Paddon, D.P., Byugen, H.S.: 2004, In *Intelligent Information Processing and Web Mining*, in press

Pap, J.M., Turmon, M., Mukhtar, S., Bogart, R., Ulrich, R., Fröhlich, C., and Wehrli, Ch.: 1997, In *Correlated Phenomena at the Sun, in the Heliosphere and in Geospace*, Proc. 31st ESLAB Symp., ESA-SP **415**, 477

Qu, M., Shih, F., Jing, J., and Wang, H.: *Solar Phys*, 2004. **222**, 137

Scherrer, P.H., et al: 1995, *Solar Phys.*, **162**, 129

Snyder, J.P.: 1993, *Flattening the Earth: Two thousand years of map projections*, The University of Chicago Press, Ltd., London

Verwichte, E.: 2003, in A. Wilson (ed) *Solar Variability as an Input to the Earth's environment, International Solar Cycle Studies Symposium*, ESA-SP **535**, 169

Vlahos, L., and Georgoulis, M.K.: 2004, *Astrophys. J.*, **603**, L61

Worden, J.R., White, O.R., and Woods, T.N.,: 1998, *Astrophys. J.*, **496**, 998

Zharkova, V., Ipson, S., Benkhalil A., and Zharkov, S.: 2005, *The Artificial Intelligence Journal*, in press