Draft version January 25, 2022
Typeset using LaTeX default style in AASTeX631

Dataset

Theodore J. Lincke,[1] Elizabeth Bradley,[1,2] Varad Deshmukh,[1] and Thomas E. Berger[3]

[1] *Department of Computer Science, University of Colorado 430 UCB, Boulder, 80309-0430 CO, USA*
[2] *Santa Fe Institute, Santa Fe, 87501 NM, USA*
[3] *Space Weather Technology, Research, and Education Center (SWx TREC), University of Colorado 429 UCB, Boulder, 80309-0429 CO, USA*

(Dated: January 25 2022)

ABSTRACT

In this paper, we propose novel features extracted from an active region (AR) using its magnetic field and continuum image representations. Our approach involves first segmenting the AR based on the salient structures on the active region such as the sunspot umbra and penumbra, the neutral line, and the background, and then deriving a set of unique physical and geometric parameters for each structure. These features isolated on individual structures more accurately represent an AR than the features derived over the entire image typically used in literature. We provide both the code to generate these features and the resultant dataset.

Keywords: Active Region — Solar Physics – Datasets – python – Image Segmentation – Solar Parameterization

1. INTRODUCTION

NASA's Solar Dynamics Observatory satellite collects the magnetogram (a measure of the magnetic field strength of the photosphere) and continuum intensity [1] of the entire solar disk (the area of the sun that faces earth) approximately every 12 minutes since 2010. Stanford's Joint Science Operations Center (JSOC) has created a pipeline that scans for active regions which are regions that have a higher than normal total magnetic field. These isolated active regions are called Space weather HMI Active Region Patches (SHARP) (1). SHARPs are organized by their distinct HMI Active Region Patch Number (HARP Number) (which is a unique number that describes that active region as it moves across the solar disk) and a timestamp that encodes when the image was taken. Therefore, it is intuitive to think about a single SHARP as a collection of active regions through time. An example data point of active region 7115 through time and 3145 through time would look like this:

```
7115
├── September 9th 2017 10:00:00 UTC
├── ...
└── September 10th 2017 10:00:00 UTC
3145
├── October 12th 2019 10:20:00 UTC
├── ...
└── October 13th 2019 23:00:00 UTC
```

The main utility for these data products (and the functionality of the new datasets described in this paper) is to forecast upcoming flares within a fixed time frame. These active regions are standalone products that are not labeled as flaring or non flaring, but the Space Weather Prediction Center (SWPC) through the National Oceanic and Atmospheric Administration (NOAA) keeps a record of the size and duration of many flares throughout time that can

thli2739@colorado.edu

[1] which is theoretically dependent on the magnetic field, hence why "sunspots" show up near high magnetic field. See (2) for an analysis on how the continuum can be expressed using the magnetic field

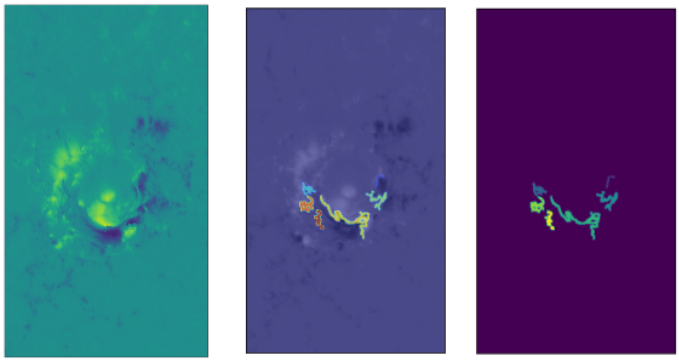


Figure 1: A neutral line shown visually. On the left is the original line of sight magnetogram; the middle shows the neutral line overlaid on top of the original magnetogram. The right image shows the isolated neutral lines (clustered based on whether pixels are touching). In the final segmentation algorithm, this Neutral Line is treated as one binary mask (not to be confused for n binary sub masks representing each neutral line).

be matched up to individual SHARPs based on the relative timing and location on the solar disk. This combination of data product (magnetogram and continuum through the SHARP) and label (flaring or non flaring through the NOAA SWPC record) forms a countable set of input output pairs with a discrete answer space. Therefore, a "flare forecasting" method (a function f_{flare}) as described in this paper will strictly be a classification problem, with a (feature extracted) SHARP as an input and a binary flaring or non flaring as an output:

$$f_{flare} : \{C \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{3 \times n \times m}\} \rightarrow \{1, 0\}$$

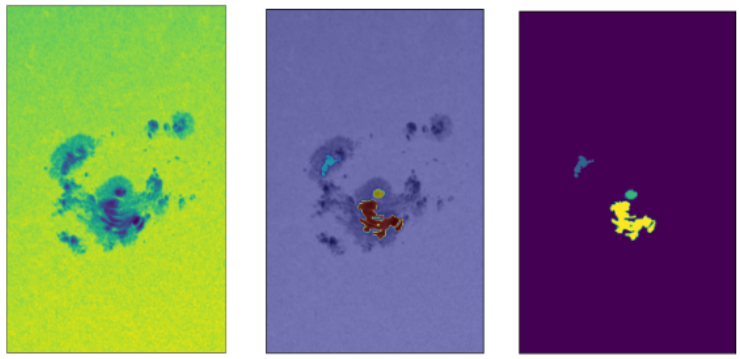
Where C is the two dimensional scalar field representing the continuum and B is the three dimensional vector field representing the magnetic field.

In this paper, these raw data products generated by JSOC are transformed into salient subregions called the Neutral Line, Umbra, Penumbra and Background. In each of these subregions, we compute a list of physical and geometric parameters (58 or more real valued scalars that describe physical properties of the active region) *restricted to the subregions*. For example, we compute the total magnetic flux *within an umbra* instead of the net magnetic flux of the entire active region.

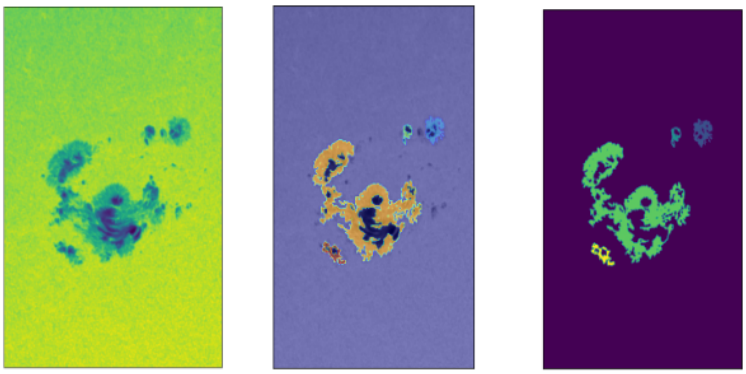
The *Neutral Line* is defined as a continuous subset of an active region such that the line of sight magnetic field is 0 and the gradient from positive to negative line of sight magnetic flux is relatively high (In this paper we use a fixed threshold and only include magnetic inversion of the active region where the absolute value of pixels is greater than $150 \frac{Mx}{cm^2}$, a value chosen by Schrijver (5)). Solar flares occur frequently in these polarity inversion regions with a high magnetic field gradient. It has been shown (4) that the length of the neutral line with horizontal shear greater than 80^o performed well in a one variable discriminant analysis. This suggests that automatically detecting the neutral lines or regions of high magnetic polarity and polarity inversion could offer more information for flare forecasting methods.

The *Umbra* and *Penumbra* are distinctly dark spots on the continuum. When the photosphere has an abnormally high magnetic field, heat is trapped beneath the surface in so-called "flux tubes" More technical term for this? - Ask Tom. The lower temperature surface ensures a lower intensity, which appears as a black or grey "sunspot" on the surface of the sun. An umbra will sometimes have an accompanying penumbra, where the continuum shows a secondary low intensity. The penumbra is seen as a visually less dark region of the active region, but still distinct from the rest of the surface. We find Umbras and Penumbras using mathematical morphology and adaptive thresholding. These distinct regions are shown visually in figure 2a and 2b.

The *background* is simply the complement of all other regions. The "Background" is not a region that has been used in solar flare forecasting. It is method specific (that is, the shape of the background will be different depending on the method used to extract the neutral line and umbra / penumbra). It is included in order to be comprehensive so we don't ignore the effects of what's happening just outside of each region. Certainly, when no other region is found (an active region has no neutral lines, umbras or penumbras), there should still be some sort of data collected on this region. Also, it is our hypothesis that the effects of the magnetogram just outside of each region will affect the shape and properties of each subregion in the future.



(a)



(b)

Figure 2: (a) An Umbra shown on an active region (shown on the left) and clustered (on the right) based on pixel intersections. (b) A penumbra shown on an active region and clustered.

The background is disjoint from all other subsets, the umbra is disjoint from the penumbra, but the neutral line is not necessarily disjoint from the umbra or penumbra.

2. METHODOLOGY

2.1. *Segmentation*

In the code provided, a single *ActiveRegion* is treated as a python object. Contained in this object are four boolean masks representing the four salient subregions described below. Each mask is a two dimensional array of the same shape as the incomming data products (the raw magnetic field, continuum). In order to generate both Baseline and Segmented datasets, we extract a set of meaningful physical features on the entire active region, then restricted to each of these masks, respectively and store the results in a comma seperated file as a list of vectors of dimension 58 and 232, respectively.

2.1.1. *Neutral Line Extraction*

The Neutral Line is defined as an area of high magnetic flux gradient and 0 magnetic flux. It is found between two large positive and negative regions in the line of sight magnetic field vector field. In order to extract the neutral line, we use a method similar to Schrijver (5). First, the line of sight magnetic field is given a threshold ($150\,\frac{Mx}{cm^2}$ is the chosen default threshold) so that any pixel between -150 and 150 is ignored. Both positive and negative subregions are dilated by 3 pixels and the intersection of these two masks is taken to be the neutral line.

Where this algorithm differs from Schrijver (5)) is the next step. In order to treat neutral lines as distinct objects, the neutral line segmentation is labeled by grouped pixels. If one group of pixels is completely disjoint from the entire netural line segmentation, then this group is called a distinct neutral line. The largest of these neutral lines is compared with all the other groups and only groups within 10% of the size of this largest neutral line are kept. All

of the portioned neutral lines are combined using a union into one large neutral line segment. The algorithm for this process is described in Algorithm 1 in the appendix.

2.1.2. *Penumbra and Umbra Extraction*

The Umbra and Penumbra are both extracted in the same algorithm. The umbra and penumbra are distinct dark subregions seen on the continuum. They are extracted first by using an adaptive localized threshold. Then, for each group of pixels (similar to the neutral lines, where each group is treated as a disjoint subset of the active region), if the difference between the maximum and minimum pixel in the continuum is greater than 21000 (from experimentation), the group of pixels then has both an umbra and penumbra. Otherwise, the group of pixels is only an umbra. For the penumbra and umbra groups, the same process is repeated and segmented once more. The final umbra and penumbra segments are filtered using the same filters for the neutral line. Ie, the largest six groups of each that are greater than 10% of the largest are kept. However, one more step is added in the filtration algorithm so that any umbra or penumbra covering more than 5% of the border of the active region is removed. This is because localized adaptive thresholding tends to favor the shadows produced at the edge of the active region and creates large groups of pixels covering a large portion of the border of the Active Region. The algorithm is described in Algorithm 2 in the appendix.

2.1.3. *Background Extraction*

The background is simply the negation of all of the previously defined subgroups.

$$Background \leftarrow \neg(Umbra \cup Penumbra \cup NeutralLine)$$

The background is included to be comprehensive. We do not wish to lose out of any information from the given active region and ignoring everything else other than the Umbra, Penumbra and Neutral Line could cause a loss of information.

2.2. *Parameters*

Recent work ((3) (4) (5)) on characterizing flaring and non flaring active regions has encouraged the use of scalar parameters based on the physical attributes of the active region. Although in any machine learning method, "more data" is always encouraged, and the pruning of this data over time to find the most optimal combination of variables speeds up and optimizes the entire forecasting pipeline, there are a few parameters that have continually been associated with high flaring rates. We have chosen a set of 58 distinct scalars that represent physical, geometric or statistical properties of subregions of an active region. For example, the total unsigned magnetic flux ($\sum_{\phi \in B_z} |\phi| dA$) is a commonly used parameter. We include all 58 features in table **??**. For any of the multiple dimensional datasets (X), we have chosen to summarize physical quantities using the first four standardized central moment: mean (μ), standard deviation ($E(X-\mu)$), skewness ($E((X-\mu)^2)$), kurtosis ($E((X-\mu)^3)$). I will refer to these four moments as $M(X)$. It should also be noted that some of these formulas require scalar multiples of physical constants, but these will be left out because every quantity is normalized.

The two datasets we extract are the Baseline, and Segmented. Baseline is a control dataset with all 58 features extracted on the entire active region and Segmented is the unique data set that extracts all 58 features for all four salient subregions from the segmentation algorithm. More precisely, an active region (AR_t^i) is a collection of pixels, each with an x, y, z magnetic flux magnitude and a continuum scalar. Therefore, one can consider the active region an element of $\mathbb{R}^{4\times n\times m}$. Let $NL_t^i, UM_t^i, PU_t^i B_t^i \subset AR_t^i$ be subsets of the active region denoting the neutral line, umbra, penumbra and background, respectively. The function $P : \mathbb{R}^4 \to \mathbb{R}^{58}$ takes in a subset of the solar surface and computes a vector of many physical features such as total magnetic flux, total shear etc. In my preliminary work, this was a vector of size 58; however, later in the ActiveRegion library, it is possible to add more features.

$$D_{\text{Baseline}} = \{P(AR_t^i) \quad : \quad \forall AR_t^i\} \tag{1}$$

The baseline dataset is simply the physical scalars of the entire active region. This is a control data set. It shows the exact same physical parameterization without any segmentation.

$$D_{\text{Segmented}} = \{\left[P\big|_{NL_t^i}(AR_t^i) \ , \ P\big|_{UM_t^i}(AR_t^i) \ , \ P\big|_{PU_t^i}(AR_t^i) \ , \ P\big|_{B_t^i}(AR_t^i)\right] \quad : \quad \forall AR_t^i\} \tag{2}$$

Where $A\big|_B$ is the restriction of function A onto the sub domain B.

When Swami is back online, I want to include an example of a few features over a time frame from the baseline and segmented data set - show the evolution of, say the total magnetic flux vs magnetic flux of the neutral line and label when the flare occurs to show that the segmented features are more indicative.

3. LIMITATIONS

Parameter tuning of the neutral line makes a difference. In the algorithm to extract the Neutral Line, a threshold value of $150\frac{Mx}{cm^2}$ is used before the positive and negative subsets are dilated and compared. This is an arbitrary number and some Active Regions require a higher or lower threshold value. A possible solution to this issue would be to adaptively estimate a threshold value based on the total magnetic flux. For example, use the 60% standard deviation of the total magnetic flux as the lower end threshold. However, this would ignore the fact that some Active Regions that may have more neutral lines, don't have a higher probability of flaring as an Active Region with a single, high flux gradient neutral line. A static flux gradient threshold is chosen to remain consistent so the measurement of the geometry and size of the neutral line is maintained.

Parameter tuning of the umbra and penumbra makes a difference. The threshold value of 21000 is chosen arbitrarily from experimentation, but removes the accuracy of a human labeled umbra and penumbra. Typically, umbras and penumbras seem obvious to the observer, but are more challenging to segment using an algorithm. Previous methods of umbra penumbra segmentation used set level image segmentation and simple statistical segmentation, but none seemed to work as well as the current proposed segmentation method using localized adaptive threshold and basic image morphology. A possible solution could include adding a pipeline that extracts umbras and penumbras using computer vision and convolutional neural networks, but this is out of the scope of this library. The library, as it stands, is designed modularly and changing the algorithm for each segmentation does not ruin the rest of the methods, so future work can still be done to tune the algorithm as it is now.

There is no global record of umbras, penumbra and neutral lines through time. This data set works well for static images for each Active Region. However, it is poorly designed for time series data. If a flare forecasting method attempts to record the physical parameterization over time, by leaving out subsets of each active region based on size, some subsets are ignored and appear / disappear over time. For example, a single Umbra that traverses the length of the Active Region could be miss identified or ignored as we take a look at later images of the same Active Region.

4. CONCLUSIONS

The proposed datasets, Baseline and Segmented, are useful feature extractions of Active Regions that can be used in flare forecasting methods. Prior parameterization from JSOC extracted only 16 physical and geometric properties. The advantages to these two datasets are two-fold. First, there are many more features extracted with the possibility of adding more. Currently, there are 58 total features, but the library is designed modularly, meaning more features can be added as a developer sees fit. The second advantage that comes only from the Segmented dataset is the restriction of physical parameters on salient subsets of the Active Region. It is widely known that much of the activity leading up to flaring comes from features of the neutral line. The complexity and almost "fractal nature" of the neutral line tends to scale with the likelihood of flaring. By restricting parameterization to salient subregions, we are better able to grasp the features of an active region. Of course, the segmentation algorithm is up to interpretation. More advanced methods for segmentation and more interesting salient subregions could exist, and the segmentation dataset, as it is now, could be further expanded on, but prior work with these two datasets shows a statistically significant difference between machine learning methods performed on the segmented and / or baseline datasets compared to the 16 physical features extracted from the SHARP's pipeline.

REFERENCES

[1]M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. D. Leka. The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: SHARPs – space-weather HMI active region patches. *Solar Physics*, 289(9):3549–3578, Sep 2014.

[2]P. Kobel, S. K. Solanki, and J. M. Borrero. The continuum intensity as a function of magnetic field. I. Active region and quiet Sun magnetic elements. *Astronomy and Astrophysics*, 531:A112, July 2011.

[3]K. D. Leka and G. Barnes. Photospheric magnetic field properties of flaring versus flare-quiet active regions. I. Data, general approach, and sample results. *The Astrophysical Journal*, 595(2):1277–1295, oct 2003.

[4]K. D. Leka and G. Barnes. Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. II. Discriminant Analysis. *The Astrophysical Journal*, 595(2):1296–1306, Oct. 2003.

[5]C. J. Schrijver. A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting. *The Astrophysical Journal*, 655(2):L117–L120, jan 2007.

APPENDIX

*

A. APPENDIX

Algorithm 1 Neutral Line Segmentation Algorithm (*ActiveRegion.assert_neutral_lines(radius (r), threshold (t))*)

Go through this algorithm and add comments for the unrelated functions and make sure it follows the format of the code. Essentially, make it better

Require: $0 < r < min(ActiveRegion.shape)$
Require: $0 < t$
  if ActiveRegion.__nl doesn't exist **then**
    $mask^+ \leftarrow$ binary_dilation(ActiveRegion.Bz $< -t$, square(r))
    $mask^- \leftarrow$ binary_dilation(ActiveRegion.Bz $> t$, square(r))
    $nl\_mask \leftarrow mask^+ \cap mask^-$
    $G \leftarrow$ group_pixels_by_touching(nl_mask)
    for $g \in G$ **do**
      if $size(g) < 10$ pixels **then**
        $G.remove(g)$
      end if
      if $size(g) < 0.1 * largest(G)$ **then**
        $G.remove(g)$
      end if
    end for
    $G \leftarrow largest\_6(G)$
    ActiveRegion.__nl $\leftarrow \cup_{\forall g \in G} g$
  end if

Algorithm 2 Umbra and Penumbra Segmentation (*ActiveRegion.assert_umbra_penumbra()*)

Go through this algorithm and add comments for the unrelated functions and make sure it follows the format of the code. Essentially, make it better. This one especially

if ActiveRegion.__umbra doesn't exist or ActiveRegion.__penumbra doesn't exist **then**
  $C_{bound} \leftarrow bound(ActiveRegion.C, 0, 255)$
  $block\_size \leftarrow min(ActiveRegion.shape)$
  if $block\_size \mod 2 = 0$ **then**
    $block\_size \leftarrow block\_size - 1$
  end if
  cont_mask $\leftarrow C_{bound} < (local\_adaptive\_threshold(C_{bound}, block\_size)$
  $G \leftarrow$ group_pixels_by_touching(cont_mask)
  for $g \in G$ **do**
    if g is touching more than 5% of the edge **then**
      $G.remove(g)$
    end if
    if $size(g) < 10$ pixels **then**
      $G.remove(g)$
    end if
    if $size(g) < 0.1 * largest(G)$ **then**
      $G.remove(g)$
    end if
  end for
  $G \leftarrow largest\_6G$
  for $g \in G$ **do**
    $t \leftarrow (max(ActiveRegion.C|_g) - min(ActiveRegion.C|_g))/2 + min(ActiveRegion.C|_g)$
    if $max(ActiveRegion.C|_g) - min(ActiveRegion.C|_g) < 21000$ **then**
      $umbra \leftarrow umbra \cup (g \cap ActiveRegion.C|_g) \leq t)$
      $penumbra \leftarrow penumbra \cup (g \cap ActiveRegion.C|_g > t)$
    else
      $umbra \leftarrow umbra \cup g$
    end if
  end for
  ActiveRegion.umbra_mask $\leftarrow umbra$
  ActiveRegion.penumbra_mask $\leftarrow penumbra$
end if

Property Description	Variable	Formula
Line of Sight Magnetic Field	$M(B_z)$	$M(B_z)$
Total of Magnetic Flux Magnitude	Φ_{tot}	$\sum_{\Phi \in B_z} \lvert\Phi\rvert dA$
Magnitude of Total Magnetic Flux	$\lvert\Phi_{net}\rvert$	$\lvert \sum_{\Phi \in B_z} \Phi dA\rvert$
Horizontal Magnetic Field	$M(B_h)$	$B_h = \lvert B_x\rvert + \lvert B_y\rvert$
Inclination Angle	$M(\gamma)$	$\gamma = arctan(\frac{B_z}{B_h})$
B Gradient	$M(\nabla B)$	$B = \lvert B_x\rvert + \lvert B_y\rvert + \lvert B_z\rvert$
B_z Gradient	$M(\nabla B_z)$	$M(\nabla B_z)$
B_h Gradient	$M(\nabla B_h)$	$M(\nabla B_h)$
Vertical Current	$M(J_z)$	$J_z = \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}$
Total of Vertical Current Magnitude	I_{tot}	$I_{tot} = \sum_{j \in J_z} \lvert j\rvert dA$
Magnitude of Total Vertical Current	$\lvert I_{net}\rvert$	$\lvert \sum_{j \in J_z} dA\rvert$
Sum of Each Polarity Current	$\lvert I^B_{net}\rvert$	$\lvert \sum_{j \in J_z (B_z > 0)} jdA\rvert + \lvert \sum_{j \in J_z (B_z < 0)} jdA\rvert$
Vertical Heterogeneity Current Density	$M(J^h_z)$	$(\frac{1}{\lvert B_x\rvert + \lvert B_y\rvert + \lvert B_z\rvert})(B_y \frac{\partial B_x}{\partial y} - B_x \frac{\partial B_y}{\partial x})$
Total of Vertical Heterogeneity Current Magnitude	I^h_{tot}	$\sum_{i \in J^h_z} \lvert idA\rvert$
Magnitude of Total Vertical Heterogeneity Current	$\lvert I^h_{net}\rvert$	$\lvert \sum_{i \in J^h_z} idA\rvert$
Twist	$M(T)$	$T = \frac{J_z}{B_z}$
Current Helicity	$M(h_c)$	$h_c = B_z(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y})$
Total of Current Helicity Magnitude	h_{ctot}	$\sum_{h \in h_c} \lvert h\rvert dA$
Magnitude of Total Current Helicity	$\lvert h_{cnet}\rvert$	$\lvert \sum_{h \in h_c} h\rvert dA$
Shear	$M(\Psi)$	$\Psi = arccos(\frac{\vec{B_p} \cdot \vec{B_o}}{B_o B_p})$
Photospheric excess Magnetic Energy Density	$M(\rho_e)$	$\lvert B_p - B_o\rvert$
Total Photospheric Excess Magnetic Energy	E_e	$\sum_{p \in p_e} dA$

Table 1: A list of all 58 physical scalars currently used by my segmentation method extracted within a subset of an active region and used as inputs to the used machine learning algorithm