Morphological Feature Extraction for Statistical Learning With Applications To Solar Image Data

David C. Stenning[1], Thomas C. M. Lee[2], David A. van Dyk[3]*, Vinay Kashyap[4], Julia Sandell[5] and C. Alex Young[6]

[1]*Department of Statistics, University of California, Irvine, CA 92617, USA*

[2]*Department of Statistics, University of California, Davis, CA 95616, USA*

[3]*Statistics Section, Department of Mathematics, Imperial College London, SW7 2AZ, UK*

[4]*High-Energy Astrophysics Division, Smithsonian Astrophysical Observatory, Cambridge, MA 02138*

[5]*Department of Physics, University of Pennsylvania, Philadelphia, PA 19104, USA*

[6]*Heliophysics Science Division, NASA/GSFC, Greenbelt, MD 20771, USA*

Received 3 May 2012; revised 21 March 2013; accepted 10 May 2013
DOI:10.1002/sam.11200
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Many areas of science are generating large volumes of digital image data. In order to take full advantage of the high-resolution and high-cadence images modern technology is producing, methods to automatically process and analyze large batches of such images are needed. This involves reducing complex images to simple representations such as binary sketches or numerical summaries that capture embedded scientific information. Using techniques derived from mathematical morphology, we demonstrate how to reduce solar images into simple 'sketch' representations and numerical summaries that can be used for statistical learning. We demonstrate our general techniques on two specific examples: classifying sunspot groups and recognizing coronal loop structures. Our methodology reproduces manual classifications at an overall rate of 90% on a set of 119 magnetogram and white light images of sunspot groups. We also show that our methodology is competitive with other automated algorithms at producing coronal loop tracings and demonstrate robustness through noise simulations. © 2013 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 6: 329–345, 2013

Keywords: mathematical morphology; image analysis; classification; sunspots; coronal loops; skeletonization

1. INTRODUCTION

The ability to extract meaningful information from large amounts of image data has far-reaching applications in such diverse fields as medicine, computer vision, and astronomy [1]. Advancements in imaging technology are yielding massive data sets that are increasingly laborious to process manually. Studying complex images 'by eye' also limits the types of analyses that can be performed since interesting features must be extracted and propagated in machine-readable form before they can be utilized in sophisticated statistical procedures. The need for automated methods is particularly apparent in the field of solar physics, with observatories such as the *Solar and Heliospheric Observatory* (SoHO), the *Transition Region and Coronal Explorer* (TRACE), and the *Solar Dynamics Observatory* (SDO) carrying high-resolution instruments operating at various wavelengths. Experts typically detect and analyze features in SoHO and TRACE images manually, but newer observatories such as SDO—with its continuous science data downlink rate of 130 Megabits per second—render impractical common labor-intensive

* *Correspondence to:* David A. van Dyk (dvandyk@imperial.ac.uk)

techniques. With routines arising from *mathematical morphology* (Section 2), we develop general techniques for extracting scientifically meaningful numerical quantities from complex high-throughput images that can be used as covariates in statistical learning methods for classification and ultimately tracking and prediction of solar features. *Our overriding goal is to extract scientifically meaningful and interpretable numerical features from solar images*. The numerical features can be carried forward into secondary analyses that will also be interpretable in terms of meaningful scientific quantities.

Our goal of extracting numerical features from images for use in secondary statistical analysis is similar in spirit to the use of functional data as predictor variables in regression. This is typically accomplished using a set of independent basis functions that represent the functional data. Although this is a mathematically attractive strategy, it does not generally lead to scientifically meaningful summaries. One notable exception involves the use of a dependent library of generating functions to represent the functional predictors [2]. This allows quantities such as the frequency, locations, and size of dips, bumps, and plateaus to be captured and passed on to the secondary analysis. Like this, we also aim to preserve scientifically meaningful summaries, but of very different predictors: images of complex solar features.

Mathematical morphology (MM) is a valuable tool for extracting shape characteristics from image data, and is well suited to the task of analyzing complex solar features. It is a nonlinear process, but we show below that it is highly effective in extracting useful numerical summaries from image data. Using appropriate morphological operations, images can be simplified by preserving the essential shape of geometric structures and eliminating noise. Therefore, MM is an excellent imaging tool for filtering, segmentation, and taking measurements such as feature areas from an image.

Our general approach to solving practical solar imaging problems is to break the original problem into a sequence of subproblems until these subproblems can be solved in a relatively simple manner. For example, one may decompose an image classification problem into the following subproblems: (i) clean the image, (ii) perform segmentation to delineate the features of interest, (iii) extract various measurements from the image, and (iv) feed these measurements to a classifier. In this example, MM can naturally be applied to solve subproblems (i) to (iii). In the remaining of this section we describe two solar imaging problems for which MM can be employed to solve some subproblems.

A major concern of current solar physics, and a stated mission for current solar observatories (http://sdo.gsfc.nasa.gov/mission/about.php), is to improve understanding of the Sun's influence on Earth and Near-Earth space. Activity in the solar *corona*—the Sun's 'atmosphere'—resulting in extreme space-weather events can have a damaging impact on Earth. In particular, highly energetic events such as *solar flares*—sudden bursts of radiation following the release of magnetic energy—and *coronal mass ejections* (CMEs)—massive bursts of coronal material—eject charged particles into space, which have the potential to damage technological infrastructure (http://www.nap.edu/catalog.php?record_id=12507). For instance, a geomagnetic storm in 1989 was responsible for the collapse of the Hydro-Québec power grid and left millions of people without power for nine hours [3]. In addition, charged particles pose a danger to astronauts on the International Space Station, or even passengers flying in aircraft at high altitude (through both exposure to radiation and the potential damage to aircraft computer systems).

Solar flares and CMEs are known to be related to various observed solar features, in particular *sunspots* and their corresponding magnetic *active regions*. Sunspots are dark areas on the Sun's *photosphere*—the region that emits the light that we see—that form when convection is inhibited by intense magnetic fields. Sunspots are classified based on the complexity of associated magnetic flux distribution as viewed in *magnetograms*, images of the spatially resolved line-of-sight magnetic field in the photosphere. One sunspot classification scheme in particular, the *Mount Wilson scheme*, has some power to predict solar flares and CMEs when combined with other space-weather data [4]. However, this classification is carried out manually and as a result is both laborious and prone to inconsistencies stemming from human observer bias [4]. That is, manual classification results in nonreproducible catalogs as two experts looking at the same set of images will not always agree. Automated sunspot classification procedures based on statistical learning methods will result in reproducible catalogs, but require numerical covariates as inputs. Using our general numerical feature extraction techniques, we produce summaries of sunspots/active regions from SoHO images that are relevant to the sunspot's classification. The scientific relevance of these numerical summaries is demonstrated by their successful use as input covariates to a supervised learning algorithm that can reproduce manual classifications with an acceptable level of agreement. As we will discuss in further detail in Section 3, it is not necessary or desirable to have the automatic classifier exactly mimic the manual class assignments. Insofar as the Mt. Wilson classification scheme contains relevant information regarding activity around a sunspot [4], by constructing numerical summaries guided by the Mt. Wilson classification rules we aim to capture the same useful scientific information. The key is that the information is obtained in a self-consistent manner, leading to more objective and reproducible data analyses. The scientific information will also be encoded in numerical feature

vectors instead of images, opening increased opportunities for downstream analyses.

Ultimately, solar physicists are interested in how features observed on the photosphere are related to volatile events originating with the release of magnetic energy in the corona. *Coronal loops*—plasma-filled structures that trace out the Sun's magnetic field—are rooted in the photosphere (the roots are referred to as footpoints) and are thus related to the morphological configurations of sunspot groups. In the vast majority of cases coronal loops are identified manually pixel-by-pixel, which is laborious and inconsistent. Hence, complex TRACE and SDO extreme ultraviolet wavelength (EUV) images provide another useful benchmark for testing our general feature extraction techniques, where the objective is to produce simple but scientifically meaningful representations of coronal loop structures that can then be used in subsequent automated procedures. Our goal is to carry out loop tracings self-consistently, based solely on the images, without invoking external factors such as magnetic field configurations. The value in these tracings is in how they are utilized in subsequent analyses by solar physicists.

The increase in quantity and quality of solar image data has spurred interest in developing automated techniques for processing such data. A general review of existing image processing techniques—including MM—useful for automated feature recognition with solar data is given in ref. 5. Simple MM is used by Curto *et al*. [6] in their procedures for automatically detecting and grouping sunspots. While this method is broadly similar to the initial step of our approach, it focuses on *identifying* sunspot groups whereas we are interested both in classifying sunspot groups and in obtaining numerical summaries of sunspot groups and active regions that can be used for statistical learning. Identification of the sunspot groups is a necessary precursor to both of these tasks. Colak and Qahwaji [7] present a system for automatically detecting and classifying sunspot groups according to the McIntosh classification scheme [8]. While we develop our methodology to match the Mt. Wilson scheme, which is more useful as a measure of the complexity of the magnetic field structure, our results and reclassifications will be applicable in either case.

Although several groups have worked on automated methods for tracing coronal loops, a satisfactory method for this challenging task remains elusive. Aschwanden *et al*. [9], for example, compare five algorithms for tracing coronal loops coming from four independent research groups and demonstrate that none of these methods can adequately reproduce results obtained from manual/visual tracing. We illustrate our method on the same test TRACE image and show that our method is competitive. Aschwanden *et al*. emphasize that comparison to manual/visual techniques is not necessarily a useful benchmark for evaluating automated routines, but the lack of robustness when comparing the various methods is disappointing. In particular, current methods for sewing together detected loop fragments and for quantifying uncertainty in traced loops are either unsatisfactory or nonexistent.

This article is divided into five sections. We begin in Section 2 with a brief introduction to MM and standard image analysis tools, and describe our general approach for extracting scientifically meaningful numerical features from images that can be used for statistical learning. In Section 3 we show how our general techniques can be used to extract numerical features from complex SoHO magnetogram and white light images that can be used in an automatic sunspot classification algorithm. In Section 4 we present an example of how our techniques can be applied to TRACE and SDO EUV images to automatically recognize and analyze coronal loops. Finally, in Section 5 we discuss our results and directions for future work. Throughout this article we use the word *feature* to describe interesting aspects of an image, such as sunspots, active regions, or coronal loops. This is not to be confused with the numerical summaries that are typically referred to as *features* in the machine learning literature. We refer to the latter as *numerical features* to avoid potential confusion. The data sets and code used to perform our analysis can be found at http://cfa.lib.harvard.edu/dvn/dv/dstenning.

2. SCIENCE-DRIVEN IMAGE ANALYSIS

The goal of science-driven image analysis is to derive scientifically meaningful quantities and machine-readable representations of images that can be used for statistical learning. MM, when combined with standard image analysis techniques, is a powerful tool for capturing the essential scientific information in a simple 'sketch' representation, a segmented image that resembles the drawing an expert would make in copying the raw image by hand. For example, a simplified representation of a coronal loop image is a binary image with pixels corresponding to the loop structure assigned a value of one. Magnetograms can be likewise segmented into simplified 'trinary' images with regions of negative magnetic polarity, positive magnetic polarity and background assigned values of two, one, and zero, respectively. The binary/trinary images sketch the solar features of interest so that numerical summaries capturing important scientific information can be calculated.

2.1. Feature Recognition

The first step in science-driven image analysis is to recognize scientifically meaningful features. For example,

we need to be able to detect sunspots, active regions, and coronal loops in solar image data as those features provide rich information about solar processes. Here we describe two typical methods used for general feature recognition and comment on their feasibility for science-driven image analysis.

Thresholding: By looking at an intensity histogram of an image, we can often determine whether the interesting features are best identified by thresholding the histogram at some particular value. Typical strategies to determine the threshold value include using the standard deviation in the histogram, using a global or a local median filter, etc. However, this method is not universally applicable because the features of interest may not be the brightest, or may exhibit variation in intensity. There is also no justification to choose one type of thresholding over another. Thus, care must be taken to ensure that the adopted threshold is not destructive to the feature we wish to study.

Background Subtraction: Background subtraction enhances the contrast of an image by making the interesting features more prominent. Typically the background is determined locally, by measuring the intensity in pixels surrounding a feature. However, for solar features, such local determination is generally not a reliable estimator of the true background. This is because (i) background pixels will be contaminated by spillover emission from the source feature and (ii) there may be overlapping features over the alleged background pixels. We therefore do not use background subtraction to detect sunspots, active regions, or coronal loops, but nevertheless carry out this operation on the TRACE and SDO images with a view toward improving the visibility of the loops. We determine the background as an average over the border of a 10×10 pixel cell, and subtract it from the average over the inner cell (a 2×2 pixel cell for TRACE and a 3×3 pixel cell for SDO) to determine the background-subtracted source intensity. We also test the sensitivity of our procedure to variation in cell size (see Section 4.3).

2.2. Mathematical Morphology

MM is a powerful tool for extracting and processing scientific information from image data because morphological operations relate directly to the shape of observed features. Here we introduce some morphological operations that are useful in extracting scientifically meaningful numerical features from images. A more detailed introduction to morphological analysis is given in the Appendix. More in depth coverage can be found in refs 10 and 11.

Dilation and Erosion: Dilation and erosion are the two fundamental operations in MM. They form a duality and they both use a structuring element (SE) Y to probe and alter the shapes of geometric structures inside an image I. The dilation of I by Y is the set of points z such that Y *hits* I when the origin of Y is placed at z. Therefore the dilation of I always enlarges I. The erosion of I by Y is defined as the set of points z such that Y *fits* wholly inside I when the origin of Y is at z. In contrary to dilation, erosion always shrinks I. For real-valued/grayscale images, the SE smoothes the three-dimensional image surface, with the height of the image surface at each pixel being equal to its intensity value.

Morphological Opening: A morphological opening operation involves, first, an erosion of the image with a SE, followed by a dilation with the same SE. Since after an erosion, only those features in the image that are morphologically similar to the SE are still present, this effectively enhances such features in the image and smoothes them from the interior. Opening also has a filtering effect: image structures that cannot completely contain the SE are removed from the image. A simple example of a morphological opening operation on a binary image is given in Fig. 1(a).

Morphological Closing: The opposite operation to opening is morphological closing, which smoothes features from their exterior. A closing operation is a dilation, followed by an erosion, which essentially smoothes out the image and fills in gaps without degrading or distorting the salient features, as would occur with normal boxcar or Gaussian smoothing. A simple example of a morphological

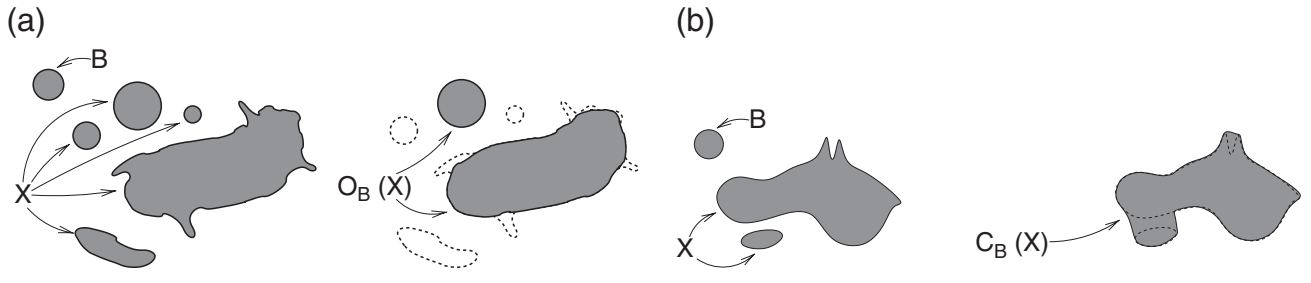

Fig. 1 Illustration of morphological opening and closing on binary images. (a) Opening of a set X by a disk B. (b) Closing of a set X by a disk B.

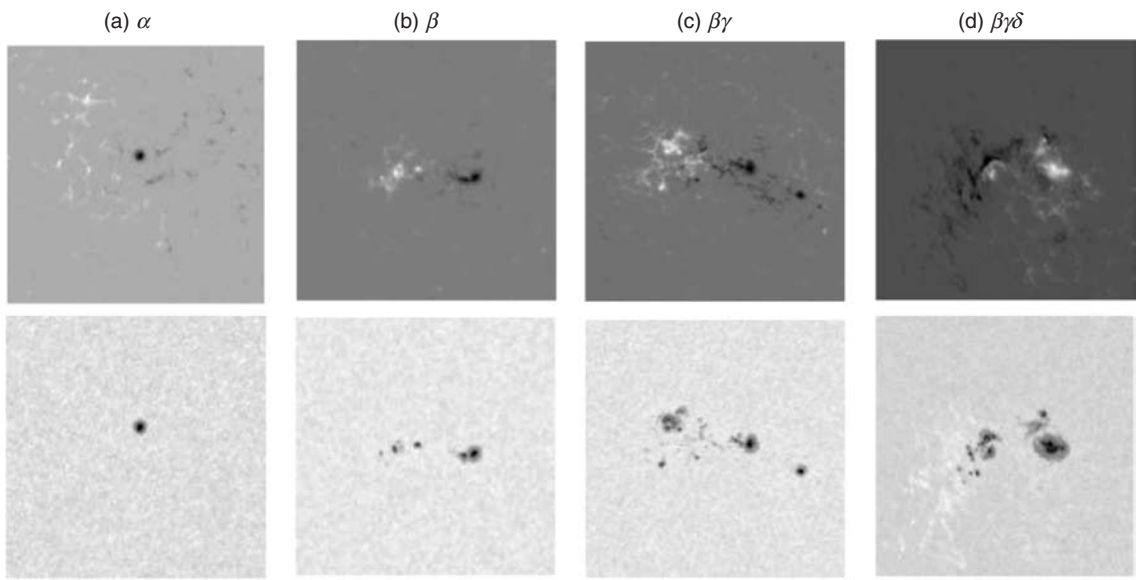

Fig. 2 Examples of the four classes of sunspot groups used in the Mt. Wilson scheme, with magnetograms in the top row and white light images in the bottom row. The α class (a) is dominated by a single unipolar sunspot that appears white or black in the magnetogram, depending on the polarity (positive or negative). The β class (b) has spots of both positive and negative polarity that can be separated by a single north–south polarity inversion line. The $\beta\gamma$ class (c) exhibits a complex distribution of polarities, and a single north–south polarity inversion line cannot cleanly divide the positive and negative regions of magnetic flux. In the $\beta\gamma\delta$ class (d), examination of the white light image in conjunction with the magnetogram reveals umbrae of different polarity within a single enclosed penumbra.

closing operation on a binary image is given in Fig. 1(b). In practice, choosing between morphological opening and closing depends on the features to be enhanced or type of noise to be removed.

Morphological Skeletonization: Skeletonization extracts the interior 'skeletons' in extended regions; the locus of the points that form the skeleton traces out the spine of the region, yielding a sketch representation of the original features. They are the innermost possible pixels in the region, and are ideally suited to capture, for example, a simplified representation of coronal loops that can then be used to extract location/shape information.

Morphological Pruning: Morphological pruning removes the small offshoots that may exist in a morphological skeleton owing to irregularities in the boundaries of the region. Such offshoots can be eliminated by first identifying the locations where the offshoots exist, then finding the lengths of such regions, and then eliminating all structures that are a few pixels long or smaller, to produce a cleaned skeleton that better represents the feature of interest.

3. SUNSPOT CLASSIFICATION

3.1. Mount Wilson Classification

The Mt. Wilson classification scheme groups sunspots into four broad classes based on the morphology of magnetically active regions as viewed in magnetogram images. Examples of the four classes appear in Fig. 2. The simplest class morphologically is the α class, defined as a single *unipolar* sunspot—a single spot of either positive or negative polarity, which is often linked to a *plage* of opposite polarity. Plage is a diffuse network of magnetic fluxtube footpoints formed when magnetic field lines shooting outward from the photosphere scatter down over a wide area. For *bipolar* sunspot groups, spots of opposite magnetic polarity are visible in magnetogram images and multiple sunspots tend to be present in the white light images, forming a *sunspot group*. The simplest bipolar class morphologically is the β class, which is a pair of sunspots of opposite magnetic polarity with a single *north–south polarity inversion line*—a simple and distinct linear spatial division oriented in the solar north–south direction—between the polarities. If a bipolar group is sufficiently complex that a single north–south polarity inversion line cannot divide the two polarities, then it is a $\beta\gamma$ sunspot group. If a $\beta\gamma$ group also contains *umbrae* of different polarity inside a single *penumbra*, which is known as a *delta spot*, then it is a $\beta\gamma\delta$ sunspot group. The umbra is the dark, inner part of the sunspot, and is surrounded by the slightly lighter penumbra as can be clearly seen in the white light image (bottom row) of Fig. 2(d).

Classification of sunspots is commonly performed through visual inspection by experts, and publicly available sunspot lists are manually determined. The Mt. Wilson

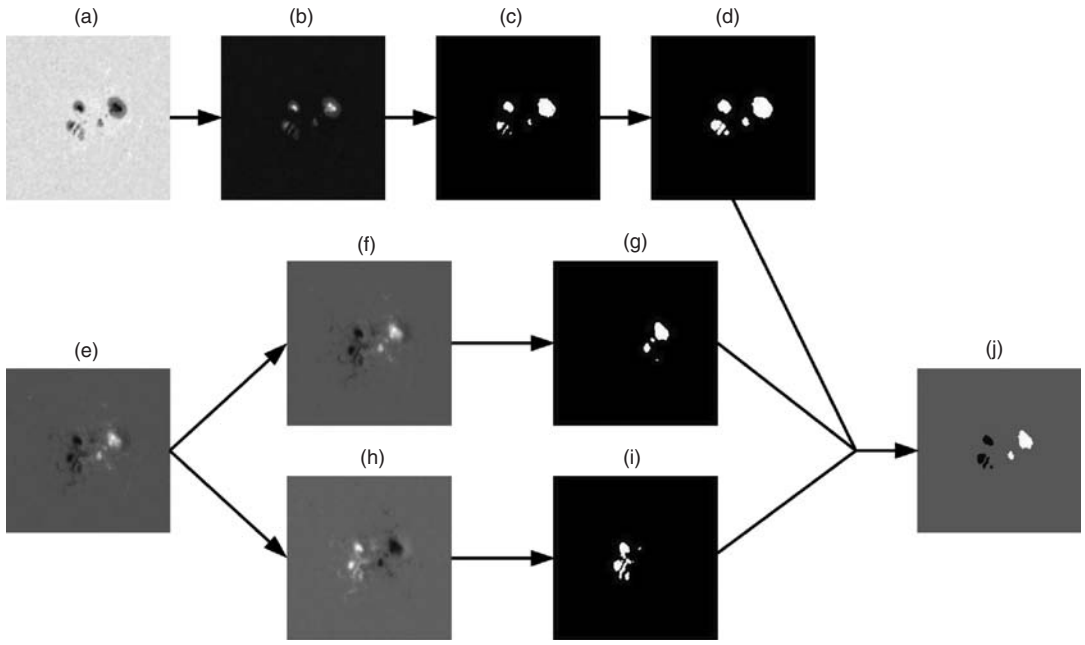

Fig. 3 Identifying active region pixels. (a) The raw β white light image. (b) The inverted white light image after applying a morphological opening operation using a spherical SE with radius 5. (c) The pixels belonging to the sunspot group identified by thresholding. (d) The sunspot area found by twice dilating the previous image using a disk-shaped SE with radius 1. (e) The raw β magnetogram. (f) The magnetogram after applying a morphological opening operation using a spherical SE with radius 1. (g) The positive polarity active region pixels identified by thresholding. (h) The inverted magnetogram after applying a morphological opening operation using a spherical SE with radius 1. (i) The negative polarity active region pixels identified by thresholding. (j) The simple active region representation found by combining the positive and negative polarity active region pixels and excluding any pixels that are not also identified as part of the sunspot area in image (d).

scheme is popular because it is based on a simple and interpretable set of rules (as described above) and has some power to predict flares when combined with other solar data [4]. However, while the classification rules are simple, the morphology of active regions is better described by a continuum rather than a discrete clustering. For example, the morphology of a particular active region may exist somewhere between a β group and a $\beta\gamma$ group and experts may disagree as to the 'correct' classification. As a result, manual classification in general suffers from human observer bias stemming from the subjective and often ambiguous morphologies of active regions [4]. A catalog of sunspot identifications and classifications constructed manually is nonreproducible, which partly motivates our automated procedure.

3.2. Generating Numerical Summaries of Solar Active Regions

In this section we describe how MM can be used to extract scientifically meaningful and statistically useful numerical features. In particular, we detail our step-by-step procedure for generating numerical summaries of active region morphology using SoHO magnetogram and white light images, improving and extending upon our work described in ref. 12. In particular, we use the white light images to obtain the general location of active regions in magnetograms to better differentiate between active region and plage network. We also calculate additional numerical summaries that characterize active region complexity that are of scientific interest in addition to serving as input covariates to statistical learning algorithms aimed at sunspot classification. Our general strategy is to obtain simple sketches (in the form of trinary images) of sunspot groups in white light images and magnetically active regions in magnetograms. Then, we calculate numerical summaries from the sketches that summarize the morphology of magnetic flux distribution that are relevant to a sunspot group's classification and can therefore be used for statistical learning. As these numerical summaries are based on the Mt. Wilson classification rules, they have a scientific basis and are interpretable to a solar physicist. In this way, we reduce complex images to real-valued numerical feature vectors that summarize the morphological characteristics of sunspot groups and associated active regions. Our general methodology is illustrated and summarized in a schematic form through Figs 3–5.

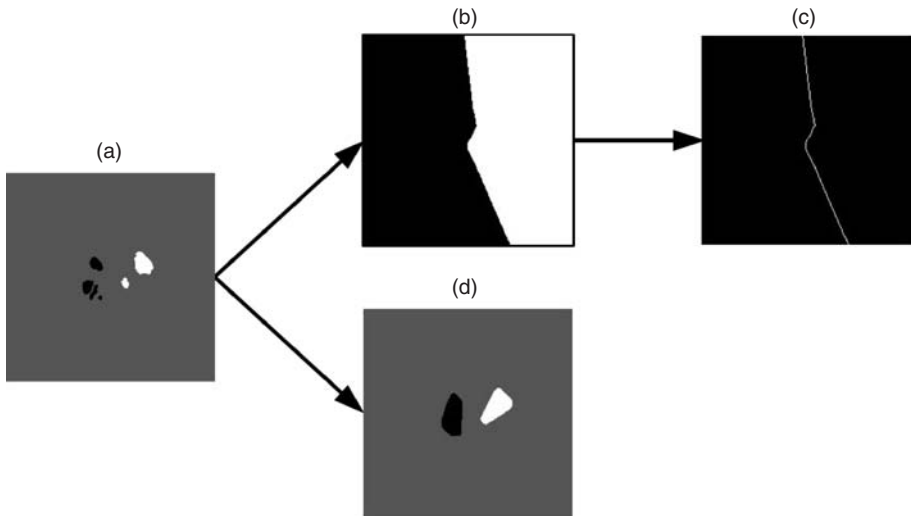

Fig. 4 Extracting numerical summaries of active regions. (a) The simple active region representation obtained through the process demonstrated in Fig. 3. (b) The separating boundary between regions of opposite magnetic polarity obtained via seeded region growing. (c) The polarity separating line obtained by removing interior and border pixels, followed by applying both a morphological opening (using a disk-shaped SE of radius one) and a morphological pruning to reduce jaggedness. (d) The simple active region representation in (a) after putting separate convex hulls around opposite polarity active region pixels.

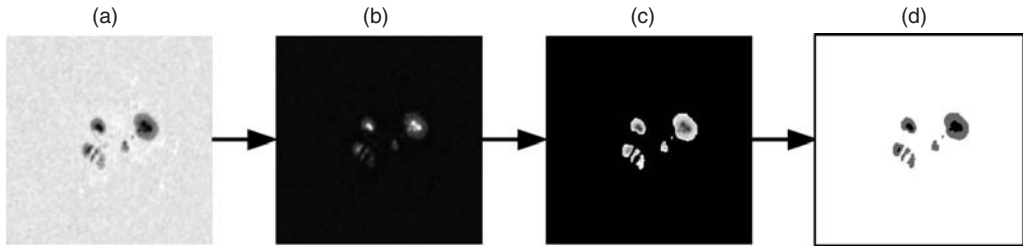

Fig. 5 Identifying delta spots. (a) The original β white light image. (b) The inverted white light image after applying a morphological opening operation using a spherical SE with radius one. (c) The pixels belonging to the sunspot group, with nonzero pixel values assigned by point-wise multiplication of the binary image obtained by thresholding image (b) and the smoothed white light image. (d) A simple representation of the umbra and penumbra regions obtained by thresholding on only the nonzero pixels from image (c). This is used in conjunction with image (j) from Fig. 3 to determine if there are umbrae with opposite polarity within a single enclosed penumbra, which is then identified as a delta spot.

3.2.1. Sunspot and active region identification

In the first part of our procedure, we use the white light images to identify sunspots. This provides us with a general location of the active regions in magnetogram images, and helps distinguish the active regions from plage network. To do this, we first clean (i.e., smooth) the inverted white light image (the image obtained by multiplying each pixel value by -1) with a morphological opening operation using a spherical SE with radius five. We are using a round SE because of the circular appearance of sunspots and active regions, and the radius was chosen so that small structures will be filtered out and larger round structures (i.e., the sunspots) will be smoothed. As at this point in our procedure we are only concerned with identifying the general location of sunspots in the white light image, we are not concerned with possible destruction of features. Next, a thresholding operation is applied by setting pixels with values above $\overline{x} + 4s$ to one and the rest to zero, where $\overline{x}$ and s are, respectively, the mean and sample standard deviation of all the pixel values in the image. The resulting binary image is dilated twice using a disk-shaped SE of radius one, which slightly increases the total area of pixels with value one. The location of sunspots in the white light image is now identified by the pixels in the binary image with value one, which we call the *sunspot area*. This process for identifying the location of sunspots is demonstrated on a β sunspot group in Fig. 3(a)-(d).

The second part of our procedure involves obtaining simple representations of active region morphology by identifying general regions of positive and negative polarity in the magnetogram. We first clean the magnetogram with a morphological opening operation using a spherical

SE of radius one. This is the smallest radius we can choose, and it is preferable since we want to remove noise while preserving and only slightly smoothing the areas of identifiable magnetic flux. We want the morphological opening to smooth the areas of positive magnetic polarity that appear white in the magnetogram so that cleaner boundaries can be obtained via thresholding (now setting pixels with values greater than $\overline{x} + 3s$ to one and the rest to zero), which identifies the *positive polarity pixels*. The same operations are applied to the inverted magnetogram (obtained by multiplying each pixel value by -1) to obtain the *negative polarity pixels*, and since we want to distinguish the two polarities these pixels are given a value of two instead of one. Any positive or negative polarity pixels (hereafter *white pixels* and *black pixels*, respectively) that fall outside the sunspot area are deemed to be part of the plage network and are set to the background value of zero. The result is a trinary sketch detailing the morphology of magnetically active regions as seen in the magnetogram. Figure 3 shows the entire process on the β active region, with arrows to illustrate how the images created at different steps of our procedure are utilized. Although the intensity information present in magnetograms is lost, this is not a detriment to our method. The type of sunspot is mainly determined by its morphological characteristics and not by the magnitude of the magnetic field. In particular, we focus on how well the magnetic flux is concentrated or scattered and how well the positive and negative polarities are mixed.

3.2.2. *Numerical features extraction*

The trinary representation of the active region, Figs 3(j) and 4(a), is our starting point for extracting numerical features that characterize the distribution of magnetic polarity. As it is obviously important to distinguish between unipolar sunspots and bipolar sunspot groups, our first numerical feature is an extreme ratio of the number of extracted white and black pixels (N_{W} and N_{B}, respectively), denoted $|N_{\mathrm{W}}/N_{\mathrm{B}}|$. The ratio is expected to be close to one for bipolar groups (β, $\beta\gamma$, and $\beta\gamma\delta$) and close to zero for α spots.

Next, we can use the amount of scatter of the white and black pixels as an indicator of active region complexity. We do this by introducing a spatial complexity measure. In particular, let W be the set of white pixels. We then compute the center of mass, c, of W. For each pixel $w \in W$, the number of pixels that a line segment from w to c passes through is denoted $L(w)$ and of these, the number of white pixels is denoted $l(w)$. The spatial complexity measure, $A(W)$, is computed as

$$A(W) = 1 - \frac{1}{|W|} \sum_{w \in W} \frac{l(w)}{L(w)},$$

where $|W|$ is the number of pixels in W. Similarly, we compute $A(B)$ for the set of black pixels. In general, we expect higher spatial complexity for $\beta\gamma$ and $\beta\gamma\delta$ sunspot groups when compared to α sunspots and β sunspot groups.

The distinction between β sunspot groups and the other bipolar sunspot groups is the presence of a single distinct north–south polarity inversion line. By treating the white and black pixels as seeds in a standard seeded region growing operation [13], we can produce a binary image that is segmented into regions of opposite polarity dominance, as in Fig. 4(b). A separating boundary for regions of opposite polarity is obtained by setting all interior and border pixels to zero, as in Fig. 4(c). A morphological opening operation using a disk-shaped SE of radius one is applied to reduce the jaggedness of the separating line, and a morphological pruning operation is applied to remove any offshoots. As sunspot groups evolve from β to $\beta\gamma$, the separating line between the two polarity regions is expected to become more complex as quantified by its curvature. We determine the amount of curvature by first tracing the separating line. The top-most pixel of the separating line is labeled as the starting point, and then we look at all the neighboring pixels to find the next pixel that is part of the separating line. This process is repeated until the entire separating line has been traced, and the separating line pixel coordinates are recorded in order as (x_i, y_i), which denotes that pixel i along the separating line is in row x and column y. Then, a quantity C_1 is computed as

$$C_1 = 1 - \frac{1}{N-1} \sum_{i=2}^{N} \frac{|x_i - x_1|}{\sqrt{(x_i - x_1)^2 + (y_i - y_1)^2}}.$$

This is a measure of curvature because the ratio inside the summation is one for a pixel lying on the 90° vertical line that intersects the starting point, so that taking the average value of this ratio for all the pixels along the separating line will yield a value of one for perfectly straight lines and smaller values as the amount of curvature increases. The average is subtracted from one so that higher values for C_1 are associated with higher curvature. We also compute C_2 in the same way as C_1, but the tracing of the separating line is initialized at its bottom-most pixel. Taking the average of C_1 and C_2 yields C, the amount of curvature of the separating line. If it happens that more than one separating line is found (as might happen with more complex sunspot groups), all the separating lines are traced and pixel coordinates are appended into a single list in order of separating line length.

The degree to which the areas of opposite magnetic polarity are 'mixing' is another useful measure of active region complexity. To determine polarity mixture, a convex hull is placed separately around the sets of white and black pixels, as in Fig. 4(d). The ratio of the number of pixels

contained inside the intersection of the two hulls, N_{in}, to the total number of pixels in the area constrained by the two hulls (where the area of overlap is only counted once), N_{out}, is a quantification of the degree of mixture, denoted by $N_{\text{in}}/N_{\text{out}}$.

The last step in our numerical feature extraction routine is to identify the presence of delta spots that distinguish $\beta\gamma$ sunspot groups from $\beta\gamma\delta$ sunspot groups. To do this, we clean the inverted white light image with a morphological operation using a spherical SE of radius one. We use a less destructive SE than in the first part of our procedure since we are concerned with distinguishing umbrae and penumbrae boundaries and do not wish to over-smooth these areas. Once we have a slightly smoothed image, we threshold to identify the *sunspot pixels*. We apply a second thresholding to the sunspot pixels to obtain *umbrae* pixels, and pixels that belong to the sunspot pixels set but not the set of umbrae pixels are designated *penumbrae pixels*. The procedure for reducing the raw white light image to a simple representation of umbra and penumbra structure is shown in Fig. 5. Because we have already identified pixels corresponding to regions of opposite magnetic polarity in the magnetogram, as in Fig. 3(j), we can examine the polarity of umbrae pixels within enclosed sets of penumbrae pixels and identify delta spots. We denote the number of delta spots detected for a particular sunspot group by N_{delta}. The size of the delta spots is also useful for distinguishing sunspots that are borderline between $\beta\gamma$ and $\beta\gamma\delta$. If delta spots are identified, then the total number of umbrae pixels associated with the delta spots is denoted S_{delta}. In cases where no delta spots are identified, both N_{delta} and S_{delta} are given values of zero.

A final caveat is that for certain unipolar sunspot groups in which only a single polarity is identified, C, $N_{\text{in}}/N_{\text{out}}$, and either $A(W)$ or $A(B)$ cannot be calculated. In these cases we assign the physically unrealistic value of negative one. We made a decision to form ratios etc., e.g., $|N_{\text{W}}/N_{\text{B}}|$ and $N_{\text{in}}/N_{\text{out}}$, prior to feeding them into a learning algorithm for classification because we want interpretable numerical features. The learning algorithm will essentially be treated as a black box that takes scientifically meaningful inputs and returns the Mt. Wilson classification.

3.3. Automatic Classification

Our numerical features, $|N_{\text{W}}/N_{\text{B}}|$, $A(W)$, $A(B)$, C, $N_{\text{in}}/N_{\text{out}}$, N_{delta}, and S_{delta}, can be used as covariates in supervised learning algorithms, using the manually determined labels. For the purposes of this article, we do not attempt to provide an optimal classifier but merely demonstrate the efficacy of our general techniques for reducing complex images into scientifically meaningful and statistically useful numerical summaries. We use Breiman's standard *random forest* [14]—a state-of-the-art nonparametric classifier that is an ensemble of individual decision trees—to exhibit how we can provide an automatic classification that is comparable to the manual classification. With N cases in the training set and p features, each tree of the random forest proceeds by sampling $n = N$ cases from the training set with replacement and randomly selecting $\sqrt{p}$ features to make a decision at each node. The desired numbers of trees are grown to completion and the classification of a new case is the majority vote of all the trees. Random forest is a sensible choice for a classifier as our features were tailored to make 'if-else' type decisions (e.g., if $|N_{\text{W}}/N_{\text{B}}| < \epsilon$ then classify as α, else continue, for a value ϵ determined by the classifier). We implement the random forest classifier using the 'randomForest' package in R. Two input variables are randomly selected as candidates for splitting at each node, and each individual tree is grown to completion. We use 1000 decision trees in total and the classifications are assigned based on majority vote.

Table 1. Confusion matrix of the random forest predictions on out-of-bag data.

		Manual classification			
		α	β	$\beta\gamma$	$\beta\gamma\delta$
Automatic classification	α	25	1	0	0
	β	2	63	5	0
	$\beta\gamma$	0	1	11	1
	$\beta\gamma\delta$	0	0	2	8

We evaluate the effectiveness of our numerical features on a dataset consisting of 119 magnetogram and white light image pairs taken by SoHO between May 1996 and December 1999. The data set was constructed by choosing magnetograms displaying individual distinct active regions that have been manually classified according to the Mt. Wilson scheme by experts at the Space Weather Prediction Center (http://www.swpc.noaa.gov/). As the training set for a particular tree in the random forest is a bootstrap sample drawn with replacement, the samples not used for that tree can be used as an out-of-bag test set for that tree. In this way, we can evaluate the random forest classifier based on the predictions on out-of-bag data as presented in Table 1. The manual and automatic classifications agree on most of the sunspot groups (the overall agreement rate is 90%), with disagreement most prominent for groups labeled manually as $\beta\gamma$. All disagreements between the automatic and manual classifiers are over adjacent classes, e.g. sunspot groups labeled manually as belonging to the $\beta\gamma$ class were placed by the automatic classifier into the β class or the $\beta\gamma\delta$ class, etc. There is a continuum of the complexity of active region morphology as we proceed from one class to the next as follows: $\alpha \rightarrow \beta \rightarrow \beta\gamma \rightarrow \beta\gamma\delta$. Because we are more concerned with capturing active region complexity in

scientifically meaningful numerical summaries than exactly mimicking the manual classification, disagreements of this type are not particularly worrisome.

The true performance of any automatic classifier will be difficult to determine since the manual classification is prone to known biases and inconsistencies. We stress that exact agreement is not necessarily the 'gold standard' when automating a manual classification that is both artificial and subjective. The true morphologies of active regions are continuous and sunspot groups can smoothly evolve from one class to another in short periods of time. As a result, there is often ambiguity as to the 'true' classification of a particular sunspot group. In Fig. 6, we present six of the sunspot groups for which there was a disagreement. The sunspot group in Fig. 6(a) was classified manually as α and automatically as β, and appears intermediate between α and β as the negative polarity is not necessarily dominant. The sunspot group in Fig. 6(b) was classified manually as β and automatically as α, but there appears to be a dominant (negative) polarity. Both panels (a) and (b) of Fig. 6 are intermediate between α and β, although Fig. 6(b) is closer to unipolar despite Fig. 6(a) having the α designation. A similar level of ambiguity exists between the sunspot groups in Fig. 6(c) and 6(d), with Fig. 6(c) having a manual classification of β and an automatic classification of $\beta\gamma$, and Fig. 6(d) having a manual classification of $\beta\gamma$ and an automatic classification of β. Both examples are intermediate between β and $\beta\gamma$. The sunspot group in Fig. 6(e) was classified manually as $\beta\gamma$ and automatically as $\beta\gamma\delta$. However, if we examine the circled areas in the magnetogram and white light images, the umbrae appear to consist of both positive and negative polarities, which would indicate the presence of a δ spot. At the very least, the presence of the δ spot is ambiguous. We can contrast this with Fig. 6(f), where the circled areas do not appear to reveal a δ spot. Since Fig. 6(f) was classified manually as $\beta\gamma\delta$ but automatically as $\beta\gamma$, this demonstrates that tuning the automatic classifier to exactly mimic the manual classification is not desirable.

Manual classification schemes must rely on an artificial discretization of active region morphologies, but an automated procedure need not be likewise hindered. With the science-driven numerical features we have extracted from magnetogram and white light images, it is possible to abandon discrete classification schemes in favor of a 'classification' based on a continuum of classes. Such an approach is expected to better capture the complex evolutionary patterns of sunspots and allow for better prediction of solar flares and CMEs. Ultimately, we wish to derive scientifically meaningful quantities from complex images that can be used to address unsolved questions in solar physics, with producing a catalogue of automatically classified sunspot groups as a by-product.

Statistical Analysis and Data Mining DOI:10.1002/sam

4. IDENTIFICATION AND ANALYSIS OF CORONAL LOOPS

4.1. Scientific Motivation

Knowledge of the coronal magnetic field is hindered by the fact that we cannot measure it directly, and therefore must carefully compare observed coronal features with predictions generated from theoretical models of the field. Specifically, we can pull out characteristics of coronal loops (loop length, temperature, etc.) to compare to predictions of different physical models. Although, as previously discussed, some progress has been made on automatically identifying coronal loops, this process typically requires manual tracing. With the high resolution of TRACE and SDO data we are able to image the coronal loops and, using methods from MM, produce automated tracings of coronal loop structures self-consistently. Solar physicists can then extract information from the coronal loop sketches that is useful for comparing and evaluating different theoretical models in an automated procedure.

In each coronal loop image, there are a large number of individual loops surrounding each active region. We must enhance dominant coronal loops from this large number of overlapping loops and also isolate the unique shape of each loop. A loop recognition mechanism consequently must be sensitive to the variations in topology and thickness, and to how close loops are to one another in order for loops to be sufficiently identified and enhanced. Currently, there are no widely implemented automated techniques for loop extraction and scientists must visually guess as to where a loop is and manually trace over the putative loops to compare them with the predictions of physical models.

To evaluate our coronal loop recognition routines, we use one image coming from TRACE and one from SDO, both obtained in the 171 Å band. The TRACE image was selected because it was used as a test image to compare five automated coronal loop tracing codes in ref. 9, so we can use those routines as a benchmark to judge the competitiveness of our MM-based methodology. We also evaluate our routine on a SDO image since SDO represents the current state of the art. The specific SDO image was chosen because of the easily discernible loop structures.

4.2. Coronal Loop Recognition

Using MM, we first experimented with different forms of rotating SEs, testing which structure detected the coronal loops with the greatest detail. A rotating rectangular SE (the axis of the rectangle is allowed to rotate to account for the various orientations of coronal loop structures) differentiates best between loops whereas a circular SE smoothes loops too greatly and loses the distinction between neighboring loops. Results are sensitive to the choice of the

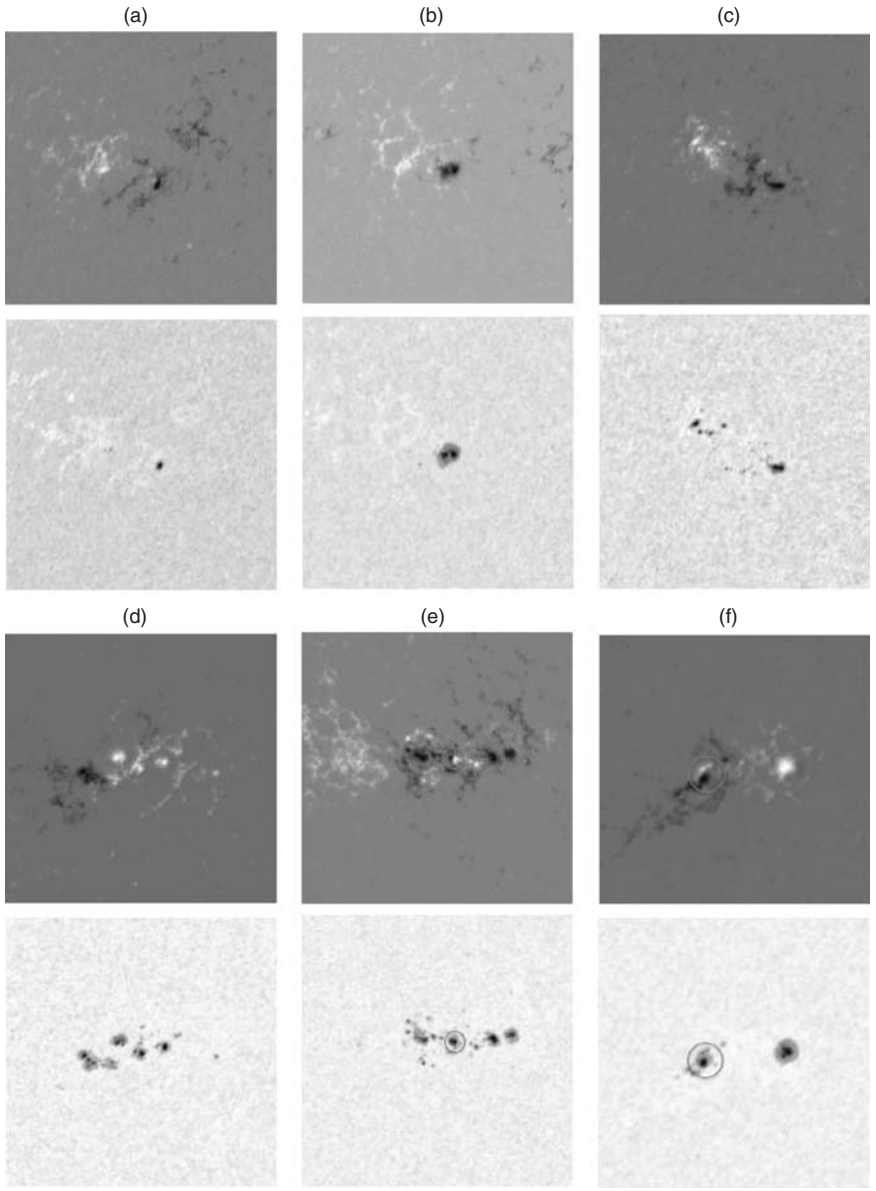

Fig. 6 Examples of the disagreements between manual and automatic classifications, with magnetograms in the top rows and white light images in the bottom rows. Classifications are given as *manual/automatic*. (a) α/β. (b) β/α. (c) $\beta/\beta\gamma$. (d) $\beta\gamma/\beta$. (e) $\beta\gamma/\beta\gamma\delta$. (f) $\beta\gamma\delta/\beta\gamma$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

angle of rotation in both opening and closing operators, as well as the size of the SEs. When closing an image, it is best to use a nonrotating rectangular SE to smooth the loop image. A rotating SE does not make loops more distinct in this case.

Upon finalizing the SE's shapes, angles of rotation, and size, a $\log_{10}$-transformation is applied to the pixel intensities. Background subtraction implemented on the coronal loop image initially and after applying the closing operator is necessary to remove components of the image which are not part of the loops. By applying the background subtraction initially we lessen the effects of lower intensity 'moss' which causes loops to appear less distinct. After subtracting the background of the image, we employ an opening operator with the rectangular rotating structure. The opening operator enhances the loop structures, and the following closing structure smoothes the enhanced loops out slightly, to increase their contour. Background subtraction is implemented again to the resulting image, further removing non-loop intensity from the image. A thresholding is applied next to increase the contrast between loops and we group pixels into isolated regions of interest

with all contiguous pixels that have values above the threshold. This process is called percolation and assigns the same label to pixels that directly neighbor each other. We use this to group pixels that comprise the detected loops together. This step is vital to extracting the loop pixels and creating a sketch representation of coronal loop structure that is necessary for comparison to physical models.

After percolation, a skeletonization function is used to reduce the loops to thin lines, revealing the distinct flux tubes of the loop structure. A pruning function is finally applied to the skeleton image, removing small 'fingers' created around the skeleton, leaving behind a trace of the coronal loop. For a visual demonstration of this process, see Fig. 7 for an example with TRACE data and Fig. 8 for an example with SDO data. Precise details on the various operations described above are given in Section 4.4.

4.3. Sensitivity Analysis

In order to test the reliability of our procedure, we carried out Monte Carlo-based simulations, varying the parameters used at each step of the process. We apply the sensitivity tests to a single identifiable loop (Fig. 9; this is a different TRACE data set than shown in Fig. 7).

We start with identifying all the pixels that form the loop in question with a basic run using a 5×5 cell for background subtraction, a rectangular SE with a width of 1 pixel and a height of 10 pixels for opening, and a square SE of size 2×2 pixels for closing. Then we perform 50 additional runs, perturbing the process at the following steps:

1. We allow for statistical variation in the data by adding a mean-zero normal random deviate at each pixel with standard deviation proportional to the square root of the intensity in that pixel, and we enforce a positivity constraint.

2. We let the background cell size vary from 5×5 to 7×7.

3. We let the rectangular SE used during morphological opening vary in width from 1 to 3 pixels.

4. We sample the height of the rectangular SE used during morphological opening from a Gaussian distribution with mean 10 and standard deviation 3.

5. We vary the magnitude of the threshold value that is dynamically determined prior to determining the contiguous regions. We do this by allowing a threshold cut to be randomly applied between the mean value of the histogram of intensities, to the mean plus three times the standard deviation observed in the histogram.

6. We choose different numbers of regions to keep in the final selection by randomly selecting from the empirically obtained distribution of the blob areas and selecting the threshold number of pixels required for a region to be kept.

At the end of each run, we obtain a skeleton of the loop of interest and all the pixels identified as part of the loop that are also in the set of loop pixels from the basic run are given a value of 1. All the images are added together, and all the pixels which are selected as part of the loop at least ten times are shown in the right panel of Fig. 9. The original loop is recovered with ease, so the process is not sensitive to the choice of input parameters used in the various operations of our procedure.

4.4. Results

Figure 10 illustrates how our MM-driven methodology can successfully trace coronal loop structure. These results are produced with the following set of input parameters: (i) The outer cell size used for background subtraction is 10×10 pixels and the inner cell size is 2×2 pixels for SDO data and 3×3 pixels for TRACE data. (ii) The rectangular SE used in opening operations has a width of 2 pixels and a height of 20 pixels. (iii) For rotation angles, we compute the opening for thirty-six separate angles, 0° to 175° in steps of 5°, and take the average of the resulting images. (iv) The morphological closing operations use a square SE of size 2×2 pixels for TRACE data and 3×3 pixels for SDO data. (v) The threshold value used in identifying contiguous regions is determined dynamically from the distribution of pixel intensities. The top 32% are left unchanged and the lower 68% are set to the value of the sixty-eighth percentile. (vi) The threshold number of pixels required to keep a region in the final skeleton is determined dynamically from the distribution of areas, with 68% of the smaller blobs discarded. (vii) The morphological skeletonization uses a circular SE with radius 1 and all 'branches' smaller than 4 pixels in length are pruned. The run time for each image is very fast, ≈ 7 seconds on a MacBook Pro.

Aschwanden *et al*. [9] compared a number of automated routines (and a manual tracing) applied to the TRACE image in Fig 7. Here we compare our techniques with the results in ref. 9, using the same quantitative criteria and notation. Figure 6 of ref. 9 presents the cumulative distribution of the number of detected coronal loops with length greater than L_{loop} pixels for five different automated

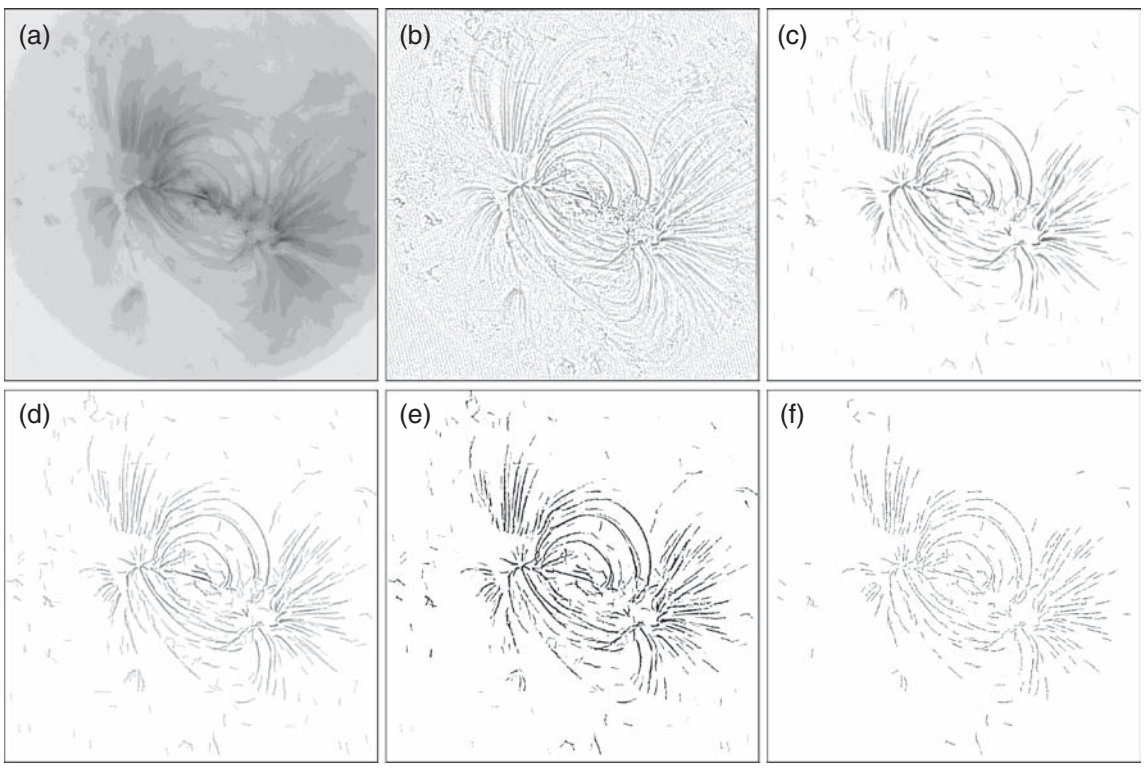

Fig. 7 The loop recognition process. (a) Original TRACE 171 Å coronal loop image (used as a test image in ref. 9 to evaluate five different automatic coronal loop tracing routines). (b) Image *a*, after applying a $\log_{10}$-transformation and subtracting the background estimated by averaging over a 10-pixel border surrounding a 2×2 central island cell. (c) Image *b*, after thresholding to exclude negative values and carrying out a morphological opening operation with a rotated, rectangular, SE. (d) Image *c*, after a morphological closing operation with a 2×2 size SE and subtracting the background again from the closed image. (e) Image *d*, after applying a thresholding to increase loop contrast and a percolation to group pixels into contiguous blobs. (f) The final coronal loop skeleton, after applying morphological skeletonization and pruning operations to Image *e*.

tracing codes and a manual tracing. We present the cumulative distribution of the coronal loop lengths detected by our automated routine in Fig. 11 for comparison. The range of maximum loop lengths reported in ref. 9 is $L_{max} =$ 244 pixels to $L_{max} = 567$ pixels, with the manual tracing yielding $L_{max} = 463$ pixels. Our automatic routine returns a maximum loop length of $L_{max} = 528$ pixels, so our result is within the range of other automated codes. Aschwanden *et al*. [9] also report that the number of detected loop segments with length $L > 70$ pixels ranges from $N(L > 70) = 30$ to $N(L > 70) = 91$ for the automated codes, with $N(L > 70) = 154$ for the manual tracing. For our routine we have $N(L > 70) = 55$, which is again within the range reported for the other automated codes. These numbers are consistent with there being no mechanism for the automated procedures to stitch together segments that the eye interprets as a continuous loop; segmentation usually arises from noise and other loops at different heights that cut across a given loop.

Demonstrating the overall significance of our results is a subtle issue since the goal is to reconstruct an image that an expert would make 'by eye' to be used in subsequent scientific analyses. The value of the coronal loop tracings is in how they are used next by solar physicists. With the skeletons we have generated it is possible to calculate physical properties of the loop structures, such as loop length and temperature at various points, and compare the empirical results to the magnetic field structure predicted by various physical models (see, for example, ref. 15). While such comparisons are not the goal of this paper, we have demonstrated how MM can be used and validated in the automatic production of coronal loop tracings useful for solar physicists.

5. DISCUSSION

The general procedures we have outlined in this paper are designed to perform science-based feature extraction. While a variety of fields are producing high-quality digital image data, most image analysis techniques rely on algorithmic operations that are not tuned to the specific

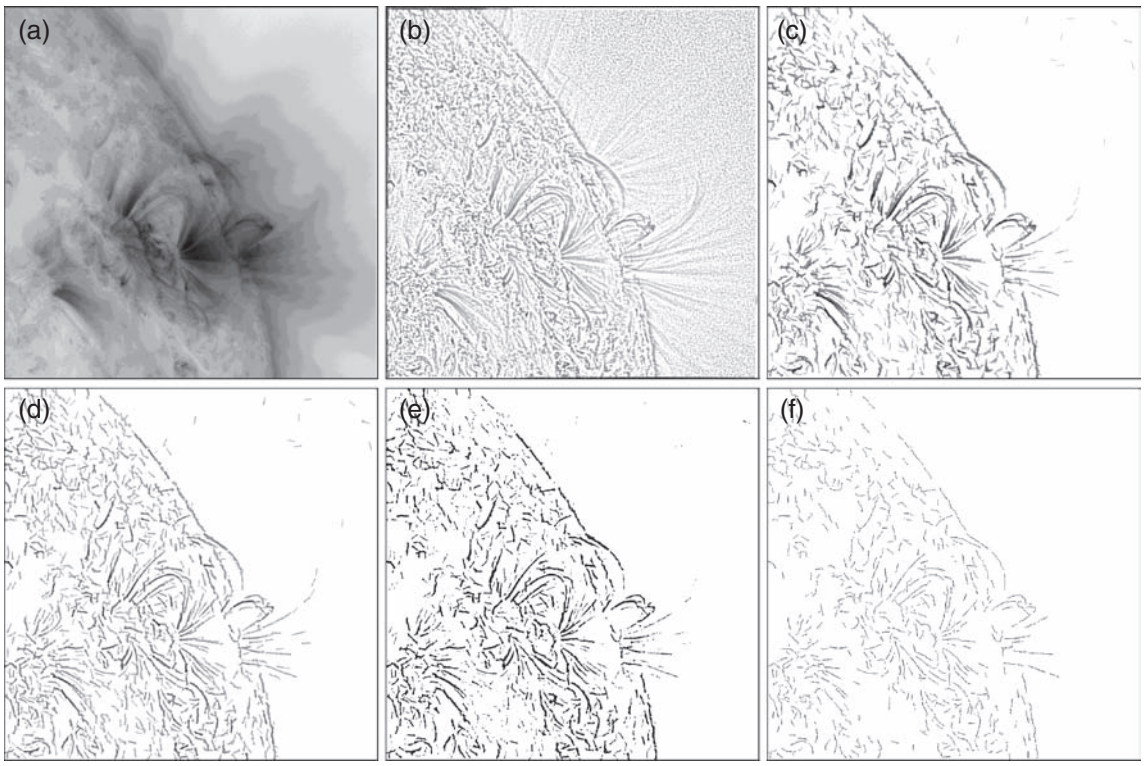

Fig. 8 The loop recognition process. (a) Original SDO 171 Å coronal loop image. (b) Image *a*, after applying a $\log_{10}$-transformation and subtracting the background estimated by averaging over a 10-pixel border surrounding a 3×3 central island cell. (c) Image *b*, after thresholding to exclude negative values and carrying out a morphological opening operation with a rotated, rectangular, SE. (d) Image *c*, after a morphological closing operation with a 3×3 size SE and subtracting the background again from the closed image. (e) Image *d*, after applying a thresholding to increase loop contrast and a percolation to group pixels into contiguous blobs. (f) Bottom/right: The final coronal loop skeleton, after applying morphological skeletonization and pruning operations to Image *e*.

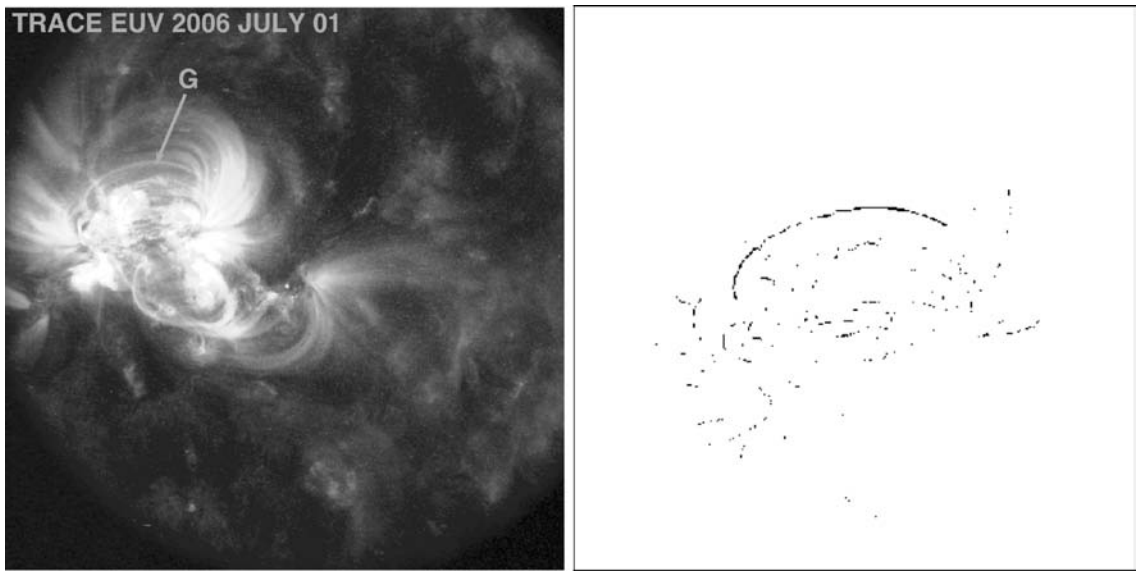

Fig. 9 The result of the sensitivity analysis. The image on the right shows the coadded skeletons of 50 runs generated using different parameters for the loop marked 'G' in the TRACE image on the left. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

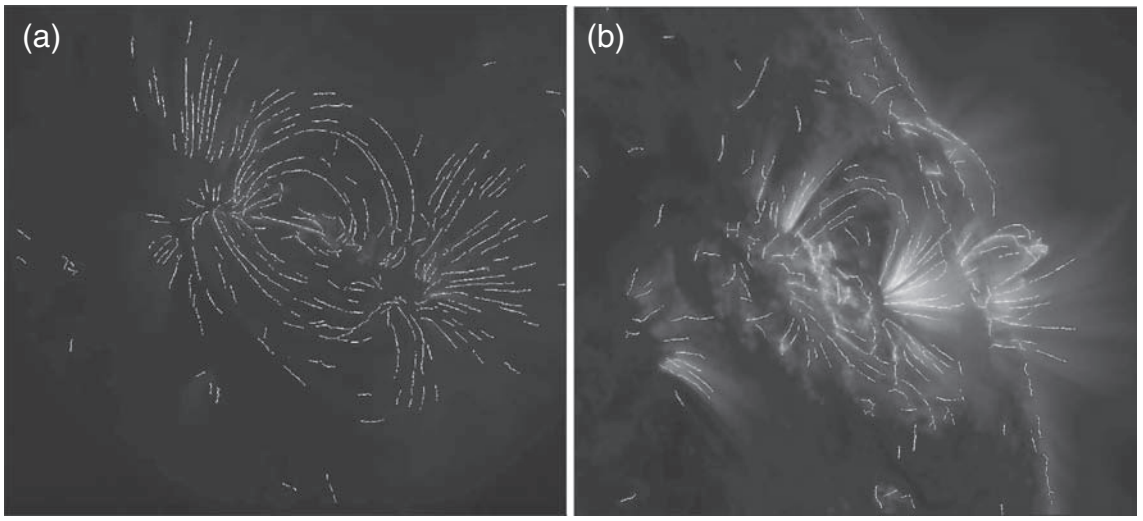

Fig. 10 Results of the automated coronal loop recognition routine. (a) Original TRACE 171 Å coronal loop image overlaid with the coronal loop tracings from Image *f* of Fig. 7. (b) Original SDO 171 Å coronal loop image overlaid with the coronal loop tracings from Image *f* of Fig. 8. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

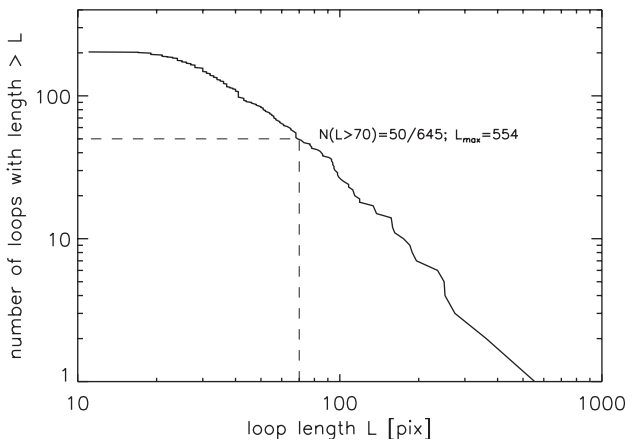

Fig. 11 The cumulative distribution of detected loop lengths —the number of detected loops with length $> L$ pixels—for our automated routine applied to the TRACE image, comparable to Figure 6 in ref. 9.

scientific information encoded in the images. We aim to purposefully extract the scientific content and use it to guide our choice of numerical features. For example, we believe there is useful information in the imperfectly performed manual sunspot classification and the Mt. Wilson scheme itself. The classes of the Mt. Wilson scheme were constructed to represent increasing complexity of magnetic field structure, but that complexity can also be captured with a set of numerical features as we demonstrated in this paper. Our automatically produced Mt. Wilson classification agrees with manually assigned classifications on 107 out of 119 images in our dataset, with disagreements occurring over adjacent classes. We also highlight that while disagreements between manual and automatic classifications will inevitably occur, it is not advantageous to tune the automatic classification routine to exactly mimic the manual classifications due to the presence of ambiguous classes and human observer bias. Insofar as the manually obtained Mt. Wilson classifications encode information about the complexities of the Sun's magnetic field, we can obtain similar information using a self-consistent and reproducible automated method.

The fact that the extracted information is now contained in scientifically meaningful numerical feature vectors is also important because it allows for further downstream analysis. An example is to use the vectors we obtain for each sunspot group in conjunction with space-weather data to determine correlations between our numerical features and solar flares. We know that the Mt. Wilson scheme has some power to predict flares when combined with other space-weather data [4], and it would be interesting if any of the numerical features we have extracted are shown to be particularly effective for predicting solar flares. One approach would be to predict flares by again feeding the numerical features into a random forest, but this time to predict whether or not a solar flare occurs in a given period of time. Random forests provide an easy method for determining which variables are the most important (see http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp for a description), so the random forest could be run again using only the variables that are deemed important. If the accuracies are similar then we have determined which numerical features (assuming there is an acceptable level of accuracy to begin with) are useful for predicting solar flares. A

researcher could also apply our methodology and compute additional numerical summaries that might be useful for different scientific investigations. An example of this might be to calculate the average gradient along the separating line in Fig. 4(c) or the maximum intensity for the umbrae pixels using Fig. 4(d) and the raw white light image. We do not extract these numerical features as they are not directly tied to the Mt. Wilson scheme, but these or similar quantities may be useful when pursuing other goals and our methodology makes them easy to calculate.

In Section 4 we demonstrated that our general heuristics can be applied to a very different type of solar imaging problem: tracing coronal loops. We demonstrated that the loop tracings can be carried out in a self-consistent manner, and the sensitivity analysis showed that the tracings are robust. The real value in the coronal loop tracings, regardless of the method in which they are obtained, is in how they are utilized by solar physicists. For instance, the loop tracings can act as a mask to pull out intensity information from the raw images that can be compared to the predictions made by theoretical models.

ACKNOWLEDGMENTS

We are grateful to three anonymous reviewers and the review editor for many helpful comments that greatly improved this paper. David van Dyk and David Stenning's research was supported in part by an NSF grant DMS-09-07522. In addition, David van Dyk was supported in part by a British Royal Society Wolfson Research Merit Award and by the STFC (UK). Vinay Kashyap acknowledges support from NASA contract NAS8-03060 to the Chandra X-ray Center. Vinay Kashyap and Julia Sandell acknowledge support from the REU Project at SAO. The work of Thomas Lee was partially supported by the National Science Foundation Grants 1007520, 1209226 and 1209232. We are grateful to Ed Deluca, Mark Weber, Monica Bobra, Aad van Ballegooijen, and Jonathan McDowell for helpful discussions.

APPENDIX

MORPHOLOGICAL OPERATIONS

Morphological operations [10,11] are set theoretic manipulations of images designed to enhance, recognize, and extract features of interest. MM provides tools that allow us to manipulate regions by filtering, thinning, pruning, etc., and describe the characteristics of regions, such as their shapes, boundaries, and skeletons. Here, we briefly describe the basics of morphological operations.

Morphological operators operate on either grayscale or binary images. Here, for simplicity and ease of explanation, we concentrate on binary images.

All morphological operations involve a SE, which is usually applied as a filter or a convolution kernel to the image. The SE is typically a simple shape, such as a 3×3 square array, or subsets thereof. Inside the morphological opening operations during our loop recognition procedure, we use slightly more complicated SEs: rectangular (1×20 to 2×70, rotatable) and annular sections (thickness of 1–2 pixels, inner radii 50–70 pixels, and subtending angles of 5–20°).

The two fundamental morphological operations are erosion and dilation:

The Erosion of an image is a set of all pixels in an image that fully contain all possible translations of the SE. In other words, for a SE Y and an image I,

$$\text{Erode}(I|Y) = \{z|Y_z \subseteq I\}\,,$$

where z are all the points such that Y, translated by z, is contained in I. Erosion shrinks an image, removing points not within in the SE and enlarging the background of an image.

The Dilation of an image is the reverse of an erosion, when the image is expanded to include all points of the SE that may overlap any point in the image. Note that in this case the SE is flipped around the origin prior to the translations, in a manner very similar to that of a convolution. Thus,

$$\text{Dilate}(I|Y) = \{z|[\hat{Y}_z \cap I] \subseteq I\}\,,$$

where $\hat{Y}$ represents Y flipped about the origin. The dilation of an image is an exact dual to an erosion of the background,

$$\text{Erode}(I|Y)^c = \text{Dilate}(I^c|\hat{Y})\,,$$

where the superscript ^{c} refers to set complement.

Sequential applications of dilation and erosion are widely used morphological operators:

Opening: An erosion, followed by a dilation, is called an Opening, i.e.

$$\text{Open}(I|S) = \text{Dilate}(\text{Erode}(I|Y))\,.$$

Thus, an opening operation ends up removing pixels from the image overall. Applying an opening operation has a smoothing effect on the image, enhancing contours and rounded structures while removing disconnected shapes.

Closing: The two operations carried out in reverse order is called a Closing, i.e.

$$\text{Close}(I|Y) = \text{Erode}(\text{Dilate}(I|Y))\,.$$

The morphological closing operation adds in pixels to smooth the image.

Skeleton: A morphological skeleton is an irreducible set of points that represent an extended region, such that each pixel on the skeleton is in some sense at the maximal distance from the edge of the region. The skeleton of an image I, given a SE Y (usually a 3×3 cell suffices),

$$\text{Skeleton}(I|Y) = \cup_{k=0}^{K}\text{Skeleton}(I|kY)\,,$$

$$\text{Skeleton}(I|kY) = \text{Erode}^k(I|Y) - \text{Open}(\text{Erode}^k(I|Y))\,,$$

where $\text{Erode}^k(I|Y)$ indicates k successive erosions of I with Y, and K is the largest value of k such that $\text{Erode}^K + 1(I|Y) = \emptyset$, a null set. Thus, the Skeleton operator acts to 'thin' an image, leaving only the spine of a region of interest.

Prune: Skeleton and other thinning operations tend to leave small offshoots that result from irregularities in the boundaries of the regions of

interest. This is usually accomplished by, first, finding all points along a skeleton that have more then two neighbors, excluding them from the image temporarily, removing all contiguous sets of pixels with fewer than some required number of pixels (typically, 4), and then reinserting the points removed previously, and dilating the image and taking its intersection with the original skeleton until no new pixels are found:

$$\begin{aligned} I_1 &= \text{convolve}(\text{Skeleton}(I|Y)|3\text{x}3) \\ I_2 &= \{x|I_1 \geq 3\} \\ I_3 &= \{x|\text{Area}(x) > 4\} \\ I_4 &= I_3 \cup I_2 \\ \text{Prune}(\text{Skeleton}(I|Y)) &= \{\text{Dilate}(I_4|3\text{x}3) \cap \text{Skeleton}(I|Y)\}^K, \end{aligned}$$

where the last operation is carried out K times such that

$$\begin{aligned} &\{\text{Dilate}(I_4|3\text{x}3) \cap \text{Skeleton}(I|Y)\}^K \\ &\quad \equiv \{\text{Dilate}(I_4|3\text{x}3) \cap \text{Skeleton}(I|Y)\}^{K+1}. \end{aligned}$$

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, Digital Image Processing (2nd ed.), Upper Saddle River, NJ, Prentice Hall, (2002).

[2] D. B. Woodard, C. Crainiceanu, and D. Ruppert, Hierarchical adaptive regression kernels for regression with functional predictors, J Comput Graph Stat (2013) DOI: 10.1080/10618600.2012.694765.

[3] L. Bolduc, GIC observations and studies in the Hydro-Québec power system, J Atmos Solar-Terr Phys 64 (16) (2002), 1793–1802.

[4] J. Ireland, C. A. Young, R. T. J. McAteer, C. Whelan, R. J. Hewett, and P. T. Gallagher, Multiresolution analysis of active region magnetic structure and its correlation with the Mt. Wilson classification and flaring activity, Solar Phys 252(1) (2008), 121–137.

[5] M. J. Aschwanden, Nonlinear force-free magnetic field fitting to coronal loops with and without stereoscopy, Astrophys J 763(2) (2013), 115.

[6] J. Curto, M. Blanca, and E. Mart'nez, Automatic sunspots detection on full-disk solar images using mathematical morphology, Solar Phys 250 (2008), 411–429. DOI: 10.1007/s11207-008-9224-6.

[7] T. Colak and R. S. R. Qahwaji, Automated mcintosh-based classification of sunspot groups using MDI images, Solar Phys 248(2) (2009), 277–296.

[8] P. S. McIntosh, The classification of sunspot groups, Solar Phys 125 (1990), 251–267.

[9] M. Aschwanden, J. Lee, G. Gary, M. Smith, and B. Inhester, Comparison of five numerical codes for automated tracing of coronal loops, Solar Phys 248 (2008), 359–377. DOI: 10.1007/s11207-007-9064-9.

[10] J. Serra. Image Analysis and Mathematical Morphology, London, New York, Academic Press, (1982).

[11] P. Soille, Morphological Image Analysis: Principles and Applications (2nd ed.), Berlin, Springer, (2003).

[12] D. Stenning, V. Kashyap, T. Lee, D. van Dyk, and C. Young, Morphological image analysis and its applications to sunspot classification, In Statistical Challenges in Modern Astronomy V, G. Jogesh Babu and Eric D. Feigelson, eds. New York, Springer, 2013, 329–342.

[13] R. Adams and L. Bischof, Seeded region growing, IEEE Trans Pattern Anal Mach Intell 16(6) (1994), 641–647.

[14] L. Breiman, Random forests, Mach Learn 45 (1) (2001), 5–32.

[15] M. Aschwanden, Image processing techniques and feature recognition in solar physics, Solar Phys 262 (2010), 235–275. DOI: 10.1007/s11207-009-9474-y.