Massive Labeled Solar Image Data Benchmarks for Automated Feature Recognition

Michael A. Schuh  
Dept of Computer Science  
Montana State University  
Bozeman, MT, 59717 USA  
michael.schuh@cs.montana.edu

Rafal A. Angryk  
Dept of Computer Science  
Georgia State University  
Atlanta, GA, 30302 USA  
angryk@cs.gsu.edu

Abstract—This paper introduces standard benchmarks for automated feature recognition using solar image data from the Solar Dynamics Observatory (SDO) mission. We combine general purpose image parameters extracted in-line from this massive data stream of images with reported solar event metadata records from automated detection modules to create a variety of event-labeled image datasets. These new large-scale datasets can be used for computer vision and machine learning benchmarks as-is, or as the starting point for further data mining research and investigations, the results of which can also aide understanding and knowledge discovery in the solar science community. Here we present an overview of the dataset creation process, including data collection, analysis, and labeling, which currently spans over two years of data and continues to grow with the ongoing mission. We then highlight two case studies to evaluate several data labeling methodologies and provide real world examples of our dataset benchmarks. Preliminary results show promising capability for the recognition of solar flare events and the classification of active and quiet regions of the Sun.

I. INTRODUCTION

The era of Big Data is here for solar physics. Capturing over 70,000 high-resolution images of the Sun per day, NASA’s Solar Dynamics Observatory (SDO) mission produces more data than all previous solar data archives combined [1]. Given this deluge of data that will likely only increase with future missions, it is infeasible to continue traditional human-based analysis and labeling of solar phenomena in every image. In response to this issue, general research in automated detection analysis is becoming increasingly popular in solar physics, utilizing algorithms from computer vision, image processing, and machine learning. The datasets presented here combine the metadata of several automated detection modules that run continuously in a dedicated data pipeline. We use these metadata catalogs to prune the massive image archive to more specific and useful forms for researchers interested in hassle-free scientific image datasets intended for region-based event (feature) recognition in each individual image over time.

This work builds upon our initial investigation into creating datasets with SDO data products [2]. We revisit the data collection process with more advanced analyses that include enhanced validation and visualization of the data streams that now extend to over two years of data starting from January 1, 2012. We then investigate, for the first time, two real world case studies of event-specific recognition tasks to provide a basic performance capability assessment for several event types with varied characteristics. The results show promising recognition capability while providing insights into several solar science-related observations and future case study possibilities.

In Section II, we provide an overview of the SDO mission, the specific data sources used, and our related previous works. Section III presents the general dataset creation process and analyses. Then two case study examples are more closely explored and evaluated in Section IV, and we briefly discuss future work and conclusions in Section V.

II. BACKGROUND

Launched on February 11, 2010, the SDO mission is the first mission of NASA’s Living With a Star (LWS) program, a long term project dedicated to studying aspects of the Sun that significantly affect human life, with the goal of eventually developing a scientific understanding sufficient for prediction [3]. The SDO is a 3-axis stabilized spacecraft in geo-synchronous orbit designed to continuously capture full-disk images of the Sun [4]. It contains three independent instruments, but our image parameter data is only from the Atmospheric Imaging

Fig. 1. An example SDO AIA image with HEK labeled event regions.

1Available at http://dmlab.cs.montana.edu/solar/
Assembly (AIA) instrument, which captures images in ten separate wavebands across the ultra-violet and extreme ultra-violet spectrum, selected to highlight specific elements of solar activity [5]. The Helioseismic and Magnetic Imager (HMI) instrument is also used to detect and characterize several types of solar events.

An international consortium of independent groups, named the SDO Feature Finding Team (FFT), was selected by NASA to produce a comprehensive set of automated feature recognition modules [1]. The SDO FFT modules operate through the SDO Event Detection System (EDS) at the Joint Science Operations Center (JSOC) of Stanford and Lockheed Martin Solar and Astrophysics Laboratory (LMSAL), as well as the Harvard-Smithsonian Center for Astrophysics (CfA), and NASA’s Goddard Space Flight Center (GSFC). Some modules are provided with direct access to the raw data pipeline for stream-like data analysis and event detection. Even though data is made publicly accessible in a timely fashion, because of the overall size, only a small window of data is available for on-demand access, while tapes provide long-term archival storage.

All events are exclusively collected from automated SDO FFT modules, removing any human-in-the-loop limitations or biases in reporting and identification. Events are reported to the Heliophysics Event Knowledgebase (HEK), which is a centralized archive of solar event reports accessible online [6]. While event metadata can be downloaded manually through the official web interface1, for efficient and automated large-scale retrieval we developed our own open source and publicly available software application named “Query HEK”, or simply QHEK4. We retrieve all event reports for seven types of solar events: active region (AR), coronal hole (CH), emerging flux (EF), filament (FI), flare (FL), sigmoid (SG), and sunspot (SS). These specific events were chosen because of their consistent and long-running modules and frequent reporting, leading to larger datasets capable of spanning longer periods of time. In the future, additional modules with similar reporting records could be included for comparison.

A summary of the event types can be found in Table I, which states the primary source (waveband or instrument) the event is reported from, the name of the reporting module, and whether or not it has a detailed chain code boundary in addition to the required bounding box outline. Also provided for comparison are the total number of event reports over the entire two years of 2012 and 2013. We note that the EF, FI, and SS events are reported from entirely different instrumentation, but we include them for the potential of novel knowledge discovery from data and because of their abundant reports and general importance to solar physics. An example AIA image with event instances overlaid on top can be seen in Fig. 1.

As another one of the 16 SDO FFT modules, our interdisciplinary research group at Montana State University (MSU) is building a “Trainable Module” for use in the first ever Content-Based Image Retrieval (CBIR) system for solar images. Each 4096 x 4096 pixel image is segmented by a fixed-size 64 x 64 grid, which creates 4096 cells per image. For each 64 x 64 pixel cell, we calculate our 10 image parameters, listed in Table II, where L stands for the number of pixels in the cell, $z_i$ is the i-th pixel value, $m$ is the mean, and $p(z_i)$ is the grayscale histogram representation of $z_i$ at $i$. The fractal dimension is calculated based on the box-counting method where $N(e)$ is the number of boxes of side length $e$ required to cover the image cell. An example 64 x 64 pixel image plot of each parameter can be seen in Fig. 2, where the colors range from blue (low values) to red (high values).

In previous work, we evaluated a variety of possible image parameters to extract from the solar images. Given the volume and velocity of the data stream, the best ten parameters were chosen based on not only their classification accuracy, but also their processing time [8], [9]. Preliminary event classification was performed on a limited set of human-labeled partial-disk images from the TRACE mission [10] to determine which image parameters best represented the phenomena [11], [12]. A later investigation of solar filament classification in H-alpha images from the Big Bear Solar Observatory (BBSO) showed similar success, even with noisy region labels and a small subset of our ten image parameters [13].

### TABLE II. THE MSU FFT IMAGE PARAMETERS.

<table>
<thead>
<tr>
<th>Label</th>
<th>Name</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Entropy</td>
<td>$E = -\sum_{i=1}^{L} p(z_i) \log_2 p(z_i)$</td>
</tr>
<tr>
<td>P2</td>
<td>Mean</td>
<td>$m = \frac{1}{L} \sum_{i=0}^{L-1} z_i$</td>
</tr>
<tr>
<td>P3</td>
<td>Standard Deviation</td>
<td>$\sigma = \sqrt{\sum_{i=0}^{L-1} (z_i - m)^2}$</td>
</tr>
<tr>
<td>P4</td>
<td>Fractal Dimensionality</td>
<td>$D_{fractal} = \log N(e) \log \epsilon$</td>
</tr>
<tr>
<td>P5</td>
<td>Skewness</td>
<td>$\mu_3 = \frac{1}{L} \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$</td>
</tr>
<tr>
<td>P6</td>
<td>Kurtosis</td>
<td>$\mu_4 = \frac{1}{L} \sum_{i=0}^{L-1} (z_i - m)^4 p(z_i)$</td>
</tr>
<tr>
<td>P7</td>
<td>Uniformity</td>
<td>$U = \frac{\sum_{i=0}^{L-1}</td>
</tr>
<tr>
<td>P8</td>
<td>Relative Smoothness</td>
<td>$R = 1 - \frac{1}{\epsilon \sum_{i=0}^{L-1}</td>
</tr>
<tr>
<td>P9</td>
<td>T. Contrast</td>
<td><em>see Tamura [7]</em></td>
</tr>
<tr>
<td>P10</td>
<td>T. Directionality</td>
<td><em>see Tamura [7]</em></td>
</tr>
</tbody>
</table>

In this work, we discuss the data collection, analysis, and transformation steps required to create our labeled datasets. As individual datasets may require different choices in a variety of methodologies throughout these steps, here we emphasize the fundamental processes and initial data source validations, with additional dataset-specific actions provided in the later case studies. We will also highlight the 5 V’s of Big Data (Volume, Velocity, Variety, Veracity, and Value) that are all exemplified in the solar data repositories presented here. Due to limited space, we focus most of our presentation on data from the single month of January 2012, with similar (and expanded)
results available on our website\(^5\) for all monthly and up-to-date cumulative statistics and datasets.

\textbf{A. Collection}

The Trainable Module parameter data is routinely mirrored from CfA to MSU, where we use it in-house for a variety of research projects and services, including our publicly available solar CBIR system\(^6\). Much work has gone into careful scrutiny of the raw data sources being collected. Some of this is out of necessity, as the entire process is highly automated and still ongoing as new data is constantly generated, so basic statistics and visualizations help inform the human maintainer. More importantly, in the era of Big Data much concern should be placed on the cleanliness of the initial data, which because of the volume is often more noisy than smaller datasets. Large-scale messy data can propagate through models and lead to less valuable (or worse yet, entirely misleading) results.

We operate in the SDO data stream at a static six minute cadence (velocity) on all ten AIA waveband channels (except 4500, which runs every 60 minutes). This totals an expected volume of 2,184 parameter files per day, and 797,160 files per year. We note again that each file contains 4,096 image cells, which results in an expected 8,945,664 cells per day and over 3.2 billion cells per year. Besides the total file and cell counts, we can visualize the time difference between each processed file for each wave, and quickly assess any large outages or cadence issues. An example for January 2012 can be seen in Fig. 3, where we plot the time difference in minutes between each file, for each separate wave. Note only four small “hiccups” are seen across all waves, and in total we have 98\% (66333 / 67704) of the expected files.

In similar fashion, we use QHEK to routinely pull all event instances from the SDO FFT modules for the seven event types previously described (see Table I). The biggest difference with this data is the veracity of multiple independent sources and the non-uniform cadence of data. Again, this can be well visualized with a time difference plot of the time between event reports for each event type. Presented in Fig. 4, instead of looking for expected smoothness in the plotted time differences (in hours), we can quickly identify the variety of reporting velocities for each event type. We also show an example highlight of a possible outage, automatically indicated if the difference between reports exceeds 24 hours (or 12 hours for our parameter data).

We can also visualize the varying volume and velocity of total event reports over time. This can be seen in Fig. 5, where we plot the total event instances reported for each unique timestamp, again over the entire month of January 2012. Note how consistent the reports and counts for AR and CH events are compared to EF or FL events. This directly conforms with the intrinsic nature of these types of events, \textit{i.e.}, active regions and coronal holes are long lasting and slowly evolving events and emerging flux and flares are relatively small and short-lived events. Also notice how this shows the steady cadence of some modules (AR, CH, SS) vs. the sporadic reporting of other modules (EF, FL, FI), which relates to the occurrence of the specific types of solar events and the data used to identify them. For example, the filament module typically only reports once or twice per day. For a broader scope, total instances for each event type over each month for the two years of data are shown in Fig. 7.

Lastly, we can verify the general cleanliness of our valuable image parameter data by visualizing a basic 3-statistic (min, max, and avg.) time-series plot of each parameter for all cells in each image. By looking at each parameter over all wavebands, we also gain a sense of how the observational source affects our parameters. For example, in Fig. 6 we show the mean parameter (P2) over one month of images. While this parameter may be less interesting than others, it is the most understandable to a human maintainer interested in sanity checking the data. Since P2 is describing the mean pixel intensity of each cell (ranging from 0 to 255), we can quickly see all the raw images are okay. Consider the obvious counter example of the 3-statistic values a solid black image would produce, which would be immediately identifiable.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{heatmap.png}
\caption{Heatmap plots of all ten image parameters for a single SDO AIA image, where each plot is normalized from 0.0 (dark blue) to 1.0 (bright red).}
\end{figure}

\(^5\)http://dmlab.cs.montana.edu/solar/

\(^6\)http://cbsir.cs.montana.edu/sdocbir
B. Transformation

All event reports are required to have a basic set of information that includes the time, location, and observation origin of the event instance. However, the characteristics of these attributes can vary greatly across different event types and reporting modules. This leads to two important issues: 1) instantaneous labeling of temporal events, 2) cross-origin event labels. For example, if an event has a lengthy duration, where and when should the report be labeled in images? And could an event label be applied across other image sources (wavebands)? Rather than attempt a singular strategy here, like previously proposed [2], we present several alternatives and considerations in our case studies. The hope is that through an accumulation of empirical results and benchmarks, we can better justify the rationale and validity of our labeling methodologies for different scenarios.
Fig. 5. The number of event reports for each unique timestamp over all event types.

Fig. 6. The 3-statistic (min, max, mean) of the parameter P2 (mean) over all AIA waveband channels.

Three spatial attributes define the event location on the solar disk. The center point and minimum bounding rectangle (MBR) are required and an optional detailed polygonal outline (chain code) may also be present. These attributes are given as geometric object strings, encapsulated by the words “POINT()” and “POLYGON()”, where point contains a single \((x, y)\) pair of pixel coordinates and polygon contains any number of pairs listed sequentially, e.g., \((x_1, y_1, x_2, y_2, \ldots, x_n, y_n)\). When the polygon is used for the MBR attribute, it always contains five vertices, where the first and last are identical. We convert all spatial attributes from the helioprojective cartesian (HPC) coordinate system to pixel-based coordinates based on imagespecific solar metadata [14], [15]. This process removes the need for any further expert knowledge or spatial positioning transformations.
C. Dataset Creation

Two important decisions impact the creation of a final dataset, 1) defining the classes to be labeled, and 2) defining data instances for such labels. Classes can represent many things, such as entire event types or subsets of event types (such as flare categories based on magnitude), while a data instance could vary from individual image cells to derived image regions to entire full-disk images. Typically we discard all image cells beyond 110% of the solar disk radius (see the larger pink ring in Fig. 8) before any region generating or class labeling is performed, as these cells are never in our scope of solar disk-based event recognition.

Through the careful combination of our cell-based image parameters and spatio-temporal event reports, we can produce labeled datasets, which are in the standard, ready-to-use, formats popular to data mining and machine learning communities. Each row of the file represents a single dataset instance (feature vector) by a simple comma-separated list of real values, with a final value representing the class label as an integer. Additionally, images (see Figs. 1 and 8) and movies can be produced to visualize the dataset instances and quickly view the large-scale datasets.

IV. CASE STUDIES

We present two case studies that serve the dual purpose of empirically showing recognition capability and comparing dataset creation decisions discussed in the previous section. Since we are focused on producing the best datasets for benchmarking automated event detection, we can use this performance measure as a possible indication of which labeling methods are more accurate. This also helps set precedence for moving forward with a general set of justified decisions for a future set of comprehensive dataset benchmarks.

Due to limited space, we again chose to present selective results highlighting only one month of data (Jan 2012). We are currently processing each individual month of data to facilitate monthly comparisons over time as well as an always up-to-date cumulative dataset benchmark. Check our website in the future for the similar analysis of other months.

A. Active Regions and Coronal Holes

We first investigate active region (AR) and coronal hole (CH) events, both of which are reported by the SPoCA module [1] in the AIA 193 Å waveband. We note the AR detection also uses AIA 171, but we chose to ignore this for a straightforward initial comparison. An example of these events are seen in Fig. 8, which also includes metadata of the solar disk center, disk radius, and our 110% radius cut-off (in pink). We can see the bounding boxes and chain codes of all event instances, as well as the highlighted cells contained within each event chain code. We also show sample “quiet” regions (QR) highlighted in gray to be used as region-based comparison in experiments.

Table III contains the total data instance counts for Jan 2012 of the various datasets presented in this work, where Remove cells are beyond our solar radius threshold and Quiet cells are non-event data instances.

<table>
<thead>
<tr>
<th>Name</th>
<th>Remove</th>
<th>Quiet</th>
<th>AR (FL)</th>
<th>CH</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCH-bbox</td>
<td>293,612</td>
<td>343,147</td>
<td>21,664</td>
<td>64,201</td>
</tr>
<tr>
<td>ARCH-ccode</td>
<td>293,612</td>
<td>383,778</td>
<td>13,676</td>
<td>31,083</td>
</tr>
<tr>
<td>ARCH-region</td>
<td>NA</td>
<td>1,227</td>
<td>1,227</td>
<td>756</td>
</tr>
<tr>
<td>FL-bbox</td>
<td>1,056,212</td>
<td>1,227</td>
<td>1,227</td>
<td>756</td>
</tr>
<tr>
<td>FL-region</td>
<td>NA</td>
<td>664</td>
<td>664</td>
<td></td>
</tr>
</tbody>
</table>

With these datasets, we can now perform additional data mining and machine learning tasks. An important example is class-based attribute analysis and evaluation. In Fig. 9 we can see the distribution of data instances for parameter P1 (Entropy) as stacked histograms for the three class labels of quiet (gray), AR (yellow), and CH (black) for the "ARCH code cells" dataset. An exhaustive set of similar plots are available for all other parameters and datasets, as well as scatterplot matrices for all two-way parameter combinations.

Since most solar phenomena occur in or near AR events, except for CH events which we also have labeled here, we can also investigate areas of the Sun not covered by either of these labels, which we call “quiet” regions. This leads to the idea of general region-based detection rather than individual cell-based detection, as well as the possibly worthwhile detection of non-event regions that are not explicitly labeled by automated modules. The third dataset we create generates a single data...
instance (feature vector) for each event label covering any number of cells. We chose to again use a simple 3-statistic summary of min, avg, and max for each original parameter over all cells in the label. Therefore, our new region-based data instances are 30 attributes long with the same class labels at the end. We can see in Table III that as one would expect, the bounding box labeled cells (ARCH-bbox) have the most data instances while the region data instances (ARCH-region) have the fewest. Quiet regions are created by replicating each AR event with an equal sized region not already labeled by any other AR or CH event. This is meant to provide a transition from cell-based to region-based event recognition.

We use several out-of-the-box machine learning algorithms on the multi-label datasets to provide a broad look at possible performance. These include: Naive Bayes (NB), decision trees (DT), support vector machines (SVM), K-nearest neighbor (KNN), and random forests (RF). Note we perform class balancing before all experiments, so the larger classes are randomly down-sampled to match the size of the smallest class. We also use all attributes and do no tuning of machine learning algorithms in these initial analyses to present a non-optimized baseline. The DT uses entropy and a max tree height of six levels, while the RF uses 10 DTs with max heights of 4 levels each. All results are the average of 10-fold cross-validation data training and testing practices.

The average accuracy of each machine learning algorithm for each variation of the ARCH datasets is presented in Fig. 11. Here we show three labeling methods for both bounding box derived label (BB) and chain code derived labels (CC). The first (ccells) evaluates each individually labeled cell, while the third (R-fvs) evaluates each region-based feature vector as a whole. Bridging the gap between these two variations, the second method (R-cells) utilizes individually labeled cells, but the cells are only contained in the identified regions. The main difference in choosing cells solely within regions is to assess if the it impacts the QR labeled data instances.

While most algorithms perform similar for all datasets, we are interested in several comparative results here. Specifically, the effects of labeling (bounding box vs chain code) and data instance (cell vs region). One would expect additional noise introduced from the possibly vague bounding box labels, and therefore decreased accuracy, but this looks more minimal than expected. Likewise, we find minimal differences between results where cells were taken explicitly from QR regions versus randomly sampled from any quiet area of the solar disk. We see a great increase in accuracy for decision tree methods (DT and RF) using region representations over individual cells. This is likely due to an easier greedy pick of initial attributes to use, while the SVM sees a decrease in performance because of the increased number of attributes – many of which may be unhelpful and adding needless complexity. It is also worth noting that the SVM was considerably slower than all other methods with even these modest (one month) dataset sizes.

**B. Flares**

The second case study we look at introduces the similar recognition task for flare (FL) events. We pick the single wave with the most FL events, which is AIA 131, and perform the same data labeling and dataset creation tasks. Since the FFT module does not report event chain codes, we use only the bounding box information. Through visual observations, flare reports represent relatively small and short-lived events, so while the bounding box is typically small (less erroneous than CH events for example), it visually applies to the entire duration (typically less than a few hours). Therefore, without obvious justification for any specific choice, and to continue with single image per event labeling, we chose to use the middle of the event duration as the labeling time.
We present results in Fig. 12 for the three FL datasets using the same labeling conventions presented in Fig. 11. Again we see SVMs do quite poorly in all situations, and KNN and RF do exceptionally well across all. Given the small bounding boxes (and therefore small amount of cells per event region), one would expect similar results for cells versus regions. While the results are much closer together than for the larger regions of the ARCH datasets, we again still see decision trees (DT and RF) perform much better on the derived region-based feature vectors. Both case studies suggest strong support for building more sophisticated region-based feature vector descriptions of the event-labeled cells.

V. CONCLUSIONS

This paper introduces the first benchmarks of event-specific data labeling for automated feature recognition in regions of SDO images. We successfully combine automated event labels with pre-computed grid-based image cell parameters to create a wide range of possible datasets. This work presented the background information about the volume and velocity of the various data sources, as well as a guided overview of the dataset collection and creation process. Through two case studies, we show initial results that indicate random forests can quite effectively recognize our solar events, with region-based feature vectors for both the ternary-labeled datasets (AR, CH, QR) and binary-labeled datasets (FL, QR) reaching average accuracy levels beyond 95%. Much future work revolves around additional and more thorough case studies to build a comprehensive set of benchmarks.

By providing ready-to-use datasets to the public, we hope to interest more researchers from various backgrounds (computer vision, machine learning, data mining, etc.) in the domain of solar physics, further bridging the gap between many interdisciplinary and mutually-beneficial research domains. In the future, we plan to extend the dataset with: (1) a longer time frame of up-to-date and labeled data, (2) more observations from other instruments on-board SDO and elsewhere, and (3) more types of events and additional event-specific attributes for extended analysis of event “sub-type” characteristics.

ACKNOWLEDGMENT

This work was supported by National Aeronautics and Space Administration (NASA) grant award No. NNX11AM13A, and by National Science Foundation (NSF) grant award No. 1443061.

REFERENCES


60