A LARGE-SCALE SOLAR IMAGE DATASET WITH LABELED EVENT REGIONS

Michael A. Schuh[†] *Rafal A. Angryk*[†] *Karthik Ganesan Pillai*[†] *Juan M. Banda*[†] *Petrus C. Martens*[‡⋆]

[†] Dept. Computer Science, Montana State University, Bozeman, MT 59717, USA
[‡] Dept. Physics, Montana State University, Bozeman, MT 59717, USA
[⋆]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA

ABSTRACT

This paper introduces a new public benchmark dataset of solar image data from the Solar Dynamics Observatory (SDO) mission. This is the first release, which contains over 15,000 images and nearly 24,000 solar events, spanning the first six months of 2012. It combines region-based event labels from six automated detection modules, ten pre-computed image parameters for each cell over a grid-based segmentation of the full resolution images, and a lower resolution version of the images for further analysis and visualization. Together, these components serve as a standardized, ready-to-use, solar image dataset for general image processing research, without requiring the necessary background knowledge to properly prepare it. We present here the fundamental dataset creation details and outline future improvements and opportunities as data collection continues for the coming years.

Index Terms— computer vision, image processing, data mining, machine learning, dataset benchmark

1. INTRODUCTION

The Solar Dynamics Observatory (SDO) mission has ushered in the era of big data for solar physics. Capturing over 70,000 high-resolution images of the Sun per day, NASA's SDO mission will produce more data than all previous solar data archives combined [1]. This overwhelming amount of data is impossible to analyze manually, image-by-image, as was common practice in earlier years. Through necessity, automated analysis is becoming the norm, utlizing algorithms from computer vision, image processing, machine learning, and more. As a result, large-scale solar data analysis is fast emerging as a novel interdisciplinary research area with ample opportunities for a wide variety of research interests.

The dataset presented here[1] combines the metadata of several automated analysis modules that run continously in a dedicated data pipeline. We use these metadata catalogs to prune the massive data archive to a more useable form for researchers interested in a hassle-free scientific image dataset intended for single-image, region-based, event recognition.

[1]Available at `http://dmlab.cs.montana.edu/solar/data/`

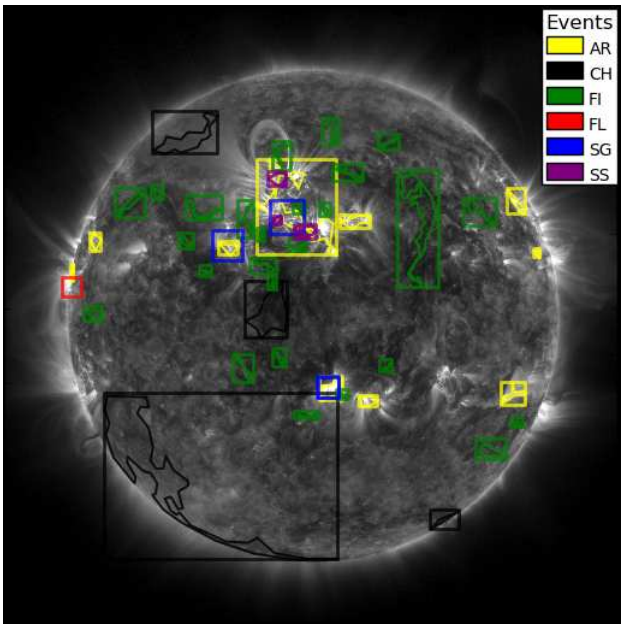

Fig. 1. An example SDO image with HEK labeled event regions.

The first release covers a six month period from January 1, 2012 to July 1, 2012, and contains over 15,000 images in two separate (but visually similar) wavebands and almost 24,000 event instances of six different event types. An example image with matching event instances is shown in Figure 1.

Along with public access and free use, a benchmark dataset such as this can offer value to potential researchers. Given the complexity of the image data and event labels, this brand new real-life dataset presents many opportunities (and challenges) for new knowledge discovery from big data in solar physics. We also pose several initial questions to explore with the dataset, motivating research in classification and clustering, image similarity indexing and retrieval, region-based object detection, and frequent pattern discovery.

It is our hope to establish an interesting, reputable, and freely-available dataset of practical use for the community. As data continues to be collected and analyzed over the coming years, we plan to release updated and expanded versions and alternative variations of the growing dataset based on research directions and community feedback.

2. BACKGROUND

Large-scale image datasets are steadily growing in availability and variety thanks to modern technology. This can be seen in all facets of life, from a personal level, with the popular crowd-sourced flickr dataset, to a national level, such as satellite imagery, and sometimes even on an international level, such as the SDO mission upon which our dataset is founded.

Benchmark datasets are crucially important to allow unbiased comparisons of independent research and development efforts across entire communities. They standardize data that is often otherwise in-exactly reproduced, causing unfavorable variabilities in published results – at best taking up valuable space in publications, and at worst altogether undocumented. Increasingly, novel research is encouraged to present on highly popular datasets supported by the community, which further and easily validates the novelty of results and claims. Exemplary datasets have hundreds of academic citations, such as the medical ImageCLEF datasets [2] and the natural scene PASCAL Visual Object Classes (VOC) datasets [3]. All of these datasets offer unique characteristics regarding the source and context of the images and labels, and this solar image dataset is no exception.

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 [4]. The SDO is a 3-axis stabilized spacecraft in geosynchronous orbit designed to continuously capture full-disk images of the Sun [5]. It contains three independent instruments (AIA, HMI, and EVE), but our dataset is currently only from the Atmospheric Imaging Assembly (AIA), which captures images in ten separate wavebands across the ultra-violet and extreme ultra-violet spectrum, selected to highlight specific elements of solar activity [6].

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[2] 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 specialized 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.

As 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 – now online and publicly available[3]. We operate at a static six minute cadence in the data pipeline on all 10 AIA wavelengths. Each 4096×4096 pixel image is segmented by a fixed-size grid, independent of any dynamic characteristics of the specific image due to long-term, real-time, stream-processing constraints. The 64×64 grid creates 4096 cells per image and our 10 image parameters (listed in Table 1) are calculated for each cell. This results in roughly 240 images (nearly one million image cells with 10 parameter values each) per 10 waves per day, or over 800,000 images per year.

Label	Name	Equation
P1	Entropy	$E = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$
P2	Mean	$m = \frac{1}{L} \sum_{i=0}^{L-1} z_i$
P3	Std. Deviation	$\sigma = \sqrt{\frac{1}{L} \sum_{i=0}^{L-1} (z_i - m)^2}$
P4	Fractal Dim.	$D_0 = \lim_{e \to 0} \frac{\log N(\epsilon)}{\log \frac{1}{\epsilon}}$
P5	Skewness	$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$
P6	Kurtosis	$\mu_4 = \sum_{i=0}^{L-1} (z_i - m)^4 p(z_i)$
P7	Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$
P8	Rel. Smoothness	$R = 1 - \frac{1}{1+\sigma^2(z)}$
P9	T. Contrast	*see Tamura [7]
P10	T. Directionality	*see Tamura [7]

Table 1. *Image Parameters, 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 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.*

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].

3. THE DATA

Here we discuss in detail the steps taken to create this dataset. While this is beneficial for reproducibility, our intent is more so to provide enough assurances to the researcher in our data curation methodologies and decisions. This also provides the reader practical knowledge of working with the several underlying large-scale data repositories.

[2] http://solar.physics.montana.edu/sol_phys/fft/

[3] http://cbsir.cs.montana.edu/sdocbir

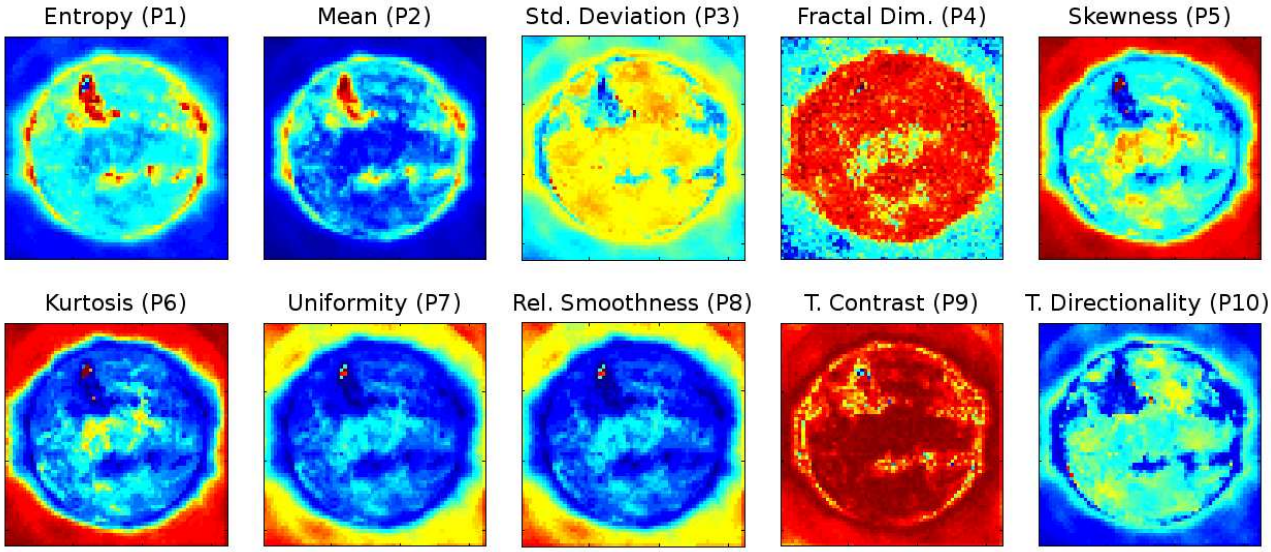

Fig. 2. Example heatmap plots of an extracted image (in 64×64 cells) for each of our ten image parameters.

3.1. Collection

All data comes from SDO FFT modules, either already available in-house at MSU or through the Heliophysics Event Knowledgebase (HEK), which is a centralized archive of metadata accessible online [14]. The HEK is an all-encompassing, cross-mission metadata repository of solar event reports and related information. This metadata can be downloaded manually through the official web interface[4], but after finding several limitations towards large-scale event retrieval, we instead developed our own open source and publicly available software application named "Query HEK", or simply QHEK[5].

We retrieved only event reports from automated SDO FFT modules for six types of solar events: active region (AR), coronal hole (CH), filament (FI), flare (FL), sigmoid (SG), and sunspot (SS). These specific events were chosen for several reasons. From a science perspective, all of these events are identifiable in static images, without requiring spatial or temporal features. These events are also, generally speaking, traditionally well-studied and frequently occuring – at least enough so that a dedicated module was created for the sole purpose of detecting such solar phenomena. From a practical perspective, this meant a larger possible set of reported event instances from well-known and well-performing automated modules that never waiver or tire, unlike graduate students.

A summary of the event types can be found in Table 2, which states the number of event instances and reported waveband. The dataset contains images in $131\mathring{A}$ and $193\mathring{A}$ wavebands that match any unique event timestamp. So the 23,517 total events represent 47,034 labels (each applied twice), but only 17,785 are true labels – AR, CH, FL, and SG events in their given wave. Note that FL and SG do not include event outlines, or chain codes (CC), and that FI and SS are reported from entirely different instrumentation. These events are included because of their abundance and importance to solar physics, and for the potential of novel knowledge discovery from data. Future dataset releases will likely include images from other wavebands and instruments.

Event	Name	Reported	CC	Instances
AR	Active Region	193 $\mathring{A}$	Yes	7108
CH	Coronal Hole	193 $\mathring{A}$	Yes	4702
FI	Filament	H-alpha	Yes	4218
FL	Flare	131 $\mathring{A}$	No	4316
SG	Sigmoid	131 $\mathring{A}$	No	1659
SS	Sunspot	HMI	Yes	1514

Table 2. *A summary of the different event types in the dataset, where CC denotes having a detailed event boundary outline.*

3.2. Transformation

We first had to standardize attributes across all event types due to independent reporting styles, such as the wavelength attribute values. For simplicity, we then discarded event instances reported in other waves, retaining the majority of total instances, but only using the two most popular waves.

Three spatial attributes define the event location on the solar disk, with the center point and minimum bounding rectangle (MBR) required. 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., $(x_1, y_1, x_2, y_2, ..., x_n, y_n)$. When the polygon is used for the MBR attribute, it always contains five

[4] `http://www.lmsal.com/isolsearch`
[5] Available at `http://dmlab.cs.montana.edu/qhek`

vertices, where the first and last are identical, while a polygonal chain code can contain any number of points. We convert all spatial attributes from the helioprojective cartesian (HPC) coordinate system to pixel-based coordinates based on image-specific metadata [15, 16]. This process removes the need for any further spatial positioning transformations.

For each event instance, we find the midpoint of the event's duration and round to the nearest minute. We then record all events that cover each unique time, and find the nearest image for each wave. This is combined to form a list of events for each unique image in each wave. We note that this is a many-to-many relationship, i.e., an image may have many assocationed events, and an event may be associated to many images. This means a single event instance might derive more than just two labels (one per wave as previously stated). Also, because we round times to the nearest minute, we buffer the event durations by ± 2 minutes so instantaneous events are not lost when matching times.

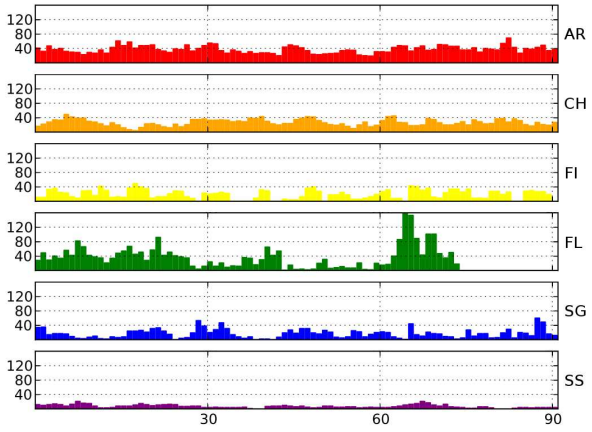

Fig. 3. Frequency of reports for all six event types over 90 days.

3.3. Formats

The dataset is released as raw text and image files. All events can be found in the *events.csv* file with the first line as the column headers. Thumbnail image and parameter data files end in *_th.png* and *.txt* respectively. Labels are provided in the *labels.txt* file, which contains one line for each image in the dataset (first value is the image file name), followed by a list of event IDs matched to the image. The entire dataset is available online [6] in a lossless compressed archive. Alternative formats available include SQL files for database useage and direct image cell feature vectors (as ARFF files for Weka [17]). Variations of the dataset will also be available, such as a class-balanced dataset for standard use by the community.

3.4. Uses

There are a variety of directions and domains this dataset can facilitate research in. At MSU, we use these data products to deliver and better develop our solar CBIR system [18]. It is also useful for benchmarking classification [13] and indexing effectiveness [19], but plenty of other interesting applications in image processing, machine learning, and data mining are possible.

Additionally, several science-related questions can now be investigated through exploratory analysis as beneficial introductory work with this dataset. For example:

- What combination of image parameters (and algorithms) work best for which types of events?
- How well can certain types of events be recognized in images they were not reported in? (e.g. FI and SS)
- Are there any event interdependencies in multi-label regions? Can this extend to spatial relationships (overlap, envelope, neighbor, etc.)?
- Quiet Sun analysis – is there really such a thing? Investigating regions of low/no event activity, much like an event type of its own.

4. CONCLUSION AND FUTURE WORK

This paper introduced the first version of a new large-scale solar image dataset, featuring full-disk images of the Sun, precomputed grid-based image cell signatures, and multi-class region-based event labels. As an extension of previous work [13], an upcoming publication using this dataset includes a detailed statistical analysis of image parameters and events, as well as basic data mining and machine learning of multi-event regions. In related works, this dataset is also being extended for use as an "event tracking" dataset, introducing dependencies on spatial and temporal attributes and exploring the possibility of mining spatio-temporal co-occurrance patterns (STCOPs) in solar physics [20, 21].

By introducing this ready-to-use dataset 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 timeframe of up-to-date and labeled data, (2) more observations from other instruments onboard SDO and elsewhere, and (3) more types of events and additional event-specific attributes for extended analysis of event "sub-type" characteristics. Further information and news about dataset updates and uses will be maintained online with the dataset for timely dissemination. We welcome and encourage the community to provide feedback about the dataset, including ideas for alternative formats and future improvements.

5. ACKNOWLEDGEMENTS

This work was supported in part by two NASA Grant Awards: 1) No. NNX09AB03G, and 2) No. NNX11AM13A.

[6] `http://dmlab.cs.montana.edu/solar/data/`

6. REFERENCES

[1] P. C. H. Martens, G. D. R. Attrill, A. R. Davey, A. Engell, S. Farid, P. C. Grigis, *et al.*, "Computer vision for the solar dynamics observatory (SDO)," *Solar Physics*, Jan 2011.

[2] W. Hersh, H. Mller, and J. Kalpathy-Cramer, "The imageclefmed medical image retrieval task test collection," *Journal of Digital Imaging*, vol. 22, pp. 648–655, 2009.

[3] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[4] G. L. Withbroe, "Living With a Star," in *AAS/Solar Physics Division Meeting #31*, vol. 32 of *Bulletin of the American Astronomical Society*, p. 839, May 2000.

[5] W. Pesnell, B. Thompson, and P. Chamberlin, "The solar dynamics observatory (sdo)," *Solar Physics*, vol. 275, pp. 3–15, 2012.

[6] J. Lemen, A. Title, D. Akin, P. Boerner, C. Chou, *et al.*, "The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO)," *Solar Physics*, vol. 275, pp. 17–40, 2012.

[7] H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 6, pp. 460–472, 1978.

[8] J. M. Banda and R. A. Angryk, "Selection of image parameters as the first step towards creating a CBIR system for the solar dynamics observatory," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 528–534, 2010.

[9] J. M. Banda and R. A. Angryk, "An experimental evaluation of popular image parameters for monochromatic solar image categorization," in *The 23rd Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 380–385, 2010.

[10] B. Handy, L. Acton, C. Kankelborg, C. Wolfson, D. Akin, *et al.*, "The transition region and coronal explorer," *Solar Physics*, vol. 187, pp. 229–260, 1999.

[11] J. M. Banda, R. A. Angryk, and P. C. H. Martens, "On the surprisingly accurate transfer of image parameters between medical and solar images," in *18th IEEE Int. Conf. on Image Processing (ICIP)*, pp. 3669–3672, 2011.

[12] J. M. Banda, R. A. Angryk, and P. C. H. Martens, "Steps toward a large-scale solar image data analysis to differentiate solar phenomena," *Solar Physics*, pp. 1–28, 2013.

[13] M. A. Schuh, J. M. Banda, P. N. Bernasconi, R. A. Angryk, and P. C. H. Martens, "A comparative evaluation of automated solar filament detection." under review, 2013.

[14] N. Hurlburt, M. Cheung, C. Schrijver, L. Chang, S. Freeland, *et al.*, "Heliophysics event knowledgebase for solar dynamics observatory (SDO) and beyond," *Solar Physics*, 2010.

[15] W. Thompson, "Coordinate systems for solar image data," *Astronomy and Astrophysics*, vol. 449, no. 2, pp. 791–803, 2006.

[16] W. D. Pence, "Cfitsio, v2.0: A new full-featured data interface," in *Astronomical Data Analysis Software and Systems*, (California), 1999.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD*, 2009.

[18] J. M. Banda, M. A. Schuh, T. Wylie, P. McInerney, and R. A. Angryk, "When too similar is bad: A practical example of the solar dynamics observatory content-based image-retrieval system," in *17th East-European Conf. on Advances in Databases and Information Systems (ADBIS)*, 2013.

[19] M. A. Schuh, T. Wylie, and R. A. Angryk, "Improving performance of high-dimensional knn retrieval through localized dataspace segmentation and hybrid indexing.," in *17th East-European Conf. on Advances in Databases and Information Systems (ADBIS)*, 2013.

[20] K. G. Pillai, R. A. Angryk, J. M. Banda, M. A. Schuh, and T. Wylie, "Spatio-temporal co-occurrence pattern mining in data sets with evolving regions," in *ICDM Workshops*, 2012, In press.

[21] K. G. Pillai, R. A. Angryk, J. M. Banda, T. Wylie, and M. A. Schuh, "Spatio-temporal co-occurrence rules," in *17th East-European Conf. on Advances in Databases and Information Systems (ADBIS)*, 2013.