# A Region-Based Retrieval System for Heliophysics Imagery

**Michael A. Schuh, Dustin Kempton, and Rafal A. Angryk**
Department of Computer Science
Georgia State University
Atlanta, GA 30302 USA
{mschuh, dkempton1, angryk}@cs.gsu.edu

**Abstract**

We introduce the creation of a new Content-Based Image Retrieval (CBIR) System for regions of interest (ROIs) in solar images. Regions are characterized by statistical features derived from general-purpose image parameters extracted in near real-time from the large-scale data stream of the Solar Dynamics Observatory (SDO) mission. This work formulates our region representation process, which includes content-based feature extraction and the derivation of various metadata features for complementary spatiotemporal similarity search capabilities. Preliminary work uses a well-established dataset of labeled event regions for supervised evaluation through event classification and retrieval performance. Feature selection is performed to reduce overall dimensionality for more effective and efficient classification and retrieval. Results show promising CBIR capabilities for region-based querying (RBQ) demands over solar image repositories.

## 1 Introduction

The Solar Dynamics Observatory (SDO) mission captures over 70,000 high-resolution images of the Sun per day, producing more data than all previous solar data archives combined (Martens et al. 2011). Given this non-stop flood of data, it has become infeasible to continue traditional human-based analysis and labeling of solar phenomena in every single image. Future endeavors will only continue to increase the volume of data, such as the currently under construction Daniel K. Inouye Solar Telescope (DKIST) that is expected to be operational by 2019 and will dwarf even the SDO data archive for decades to come (Rimmele et al. 2015). In response to this new era of Big Data, interdisciplinary research is becoming increasingly popular between computer science and solar physics, utilizing algorithms from computer vision and image processing, data mining and machine learning, and information retrieval fields.

This work builds upon existing investigations (Schuh et al. 2013) into creating machine-usable datasets with SDO data products to pursue data mining and knowledge discovery from data (KDD) efforts. The data used here combines openly available data products from several automated detection modules that run continuously in a dedicated data processing pipeline[1]. Here we use these datasets to present proof-of-concept region-based event recognition and retrieval in each individual image over time. This also establishes a baseline benchmark of performance capabilities for comparative evaluation of more advanced future research with similar objectives. Therefore, our primary focus is proper data preparation and experimental methodology for the framework as a whole.

In Section 2, we provide an overview of the SDO mission, the specific data sources used and related works. Section 3 presents the general dataset creation and region representation process. Then in Section 4, we highlight our evaluation of supervised feature selection, classification, and information retrieval. Lastly, Section V finishes with a brief discussion of future work and conclusions.

## 2 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 (Withbroe 2000). The SDO is a 3-axis stabilized spacecraft in geo-synchronous orbit designed to continuously capture images of the entire (full-disk) Sun (Pesnell, Thompson, and Chamberlin 2012). It contains three independent instruments, but our main focus is the Atmospheric Imaging 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 (Lemen et al. 2012). Figure 1 shows an example AIA 171 Ångström (Å) image. While all of these images are grayscale, they are often colorized in unique ways to quickly identify the wavelength and better accentuate the solar activity of interest.

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 (event) recognition modules (Martens et al. 2011). 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 As-



[1]More info at http://dmlab.cs.gsu.edu/solar/

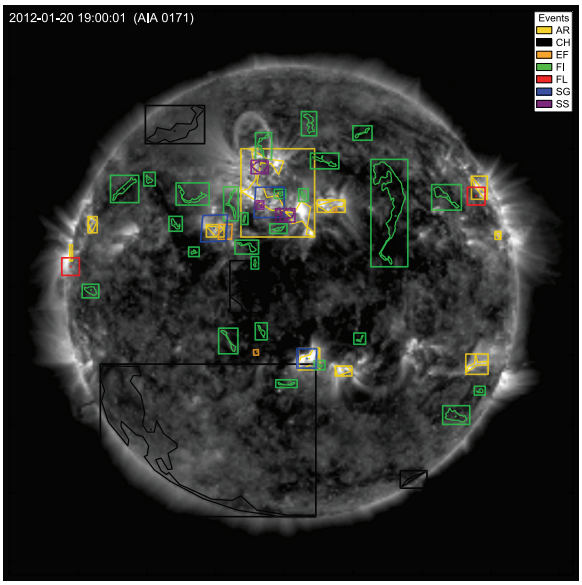


Figure 1: An example SDO AIA 171Å image overlaid with labeled event regions of interest from SDO FFT modules.

trophysics Laboratory (LMSAL), as well as the Harvard-Smithsonian Center for Astrophysics (CfA), and NASA's Goddard Space Flight Center (GSFC). Several modules are provided with direct access to the raw data pipeline for stream processing and near-real-time event detection.

Solar events identified by these modules, among others, are reported to the Heliophysics Event Knowledgebase (HEK), which is a centralized archive of solar event reports accessible online (Hurlburt et al. 2012). While event metadata can be downloaded manually through the official web interface[2], this process is cumbersome and slow. In previous work, automated methods were used to comprehensively collect and process all events reports into a clean and ready-to-use dataset. Importantly, the dataset contains events only from the automated SDO FFT modules, removing any human-in-the-loop limitations or biases in reporting and identification, while facilitating timely updates as modules continually deliver new reports. Here we begin with this curated dataset as a starting point, and refer the reader to the previous work for more detailed information (Schuh, Angryk, and Martens 2016).

Figure 1 shows examples of all seven types of events in the dataset: Active Region (AR), Coronal Hole (CH), Emerging Flux (EF), Filament (FI), Flare (FL), Sigmoid (SG), and Sunspot (SS), which were all chosen because of their long-running modules and frequent reporting. Many of these event types are identified in alternative image sources, including non-SDO data, so this figure is only meant as a general spatiotemporal reference. This work focuses on AR and CH event types. Active regions are brighter areas of higher magnetic activity that contain other events such as sunspots and flares. Conversely, coronal holes are much cooler and darker areas of lower energy and activity. Both types of events are identified by the SPoCA module (Verbeeck, C. et al. 2014) using AIA 171Å and 193Å images.

As another one of the 16 SDO FFT modules, the interdisciplinary research group at Georgia State University (GSU) developed a "Trainable Module" to comprehensively extract image parameters over each AIA image, enabling novel research efforts in a wide variety of data mining topics unrelated to specific solar phenomena, including a Content-Based Image Retrieval (CBIR) system for full-disk solar images (Banda et al. 2013) and solar event tracking (Kempton, Schuh, and Angryk 2016). The main benefit of such a module is the availability of a general-purpose and space-efficient image descriptor catalog over the entire data archive for future research that has yet to be determined.

In previous works, a wide variety of possible image parameters were evaluated to represent the solar images (Banda and Angryk 2010a; 2010b). 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. Figure 2 shows heatmap plots of each normalized parameter for a single image (very similar to the image presented in Fig. 1), where the colors range from blue (low values) to red (high values). Each $4096 \times 4096$ pixel image is segmented by a fixed-size $64 \times 64$ grid, which creates 4096 non-overlapping cells per image. For each $64 \times 64$ pixel cell, these ten image parameters are calculated and archived along with image metadata and thumbnails.

The Trainable Module operates on a six minute cadence that results in roughly 240 images processed per day for each of the ten SDO AIA channels (wavebands), which is over 800,000 images per year. This totals approximately 850MB per day, which sums to about 300GB per year. Much like the event datasets, a previous work has meticulously analyzed and curated large-scale datasets to make this image parameter data easily available for research such as this work, and we refer the reader here for more detailed information beyond our scope (Schuh, Angryk, and Martens 2015).

## 3 The Data

Here we discuss the creation of our specific CBIR evaluation dataset using the public data repositories discussed in the previous section. We first provide an overview of the region identification process, which links event labels with associated spatiotemporal regions of image parameter data. Then using this set of labeled regions, we discuss our specific representation (or characterization) methodology, which forms the labeled data instances for our research dataset. We emphasize the representation process as an independent step after identification, because the larger goal is that alternative representations, such as sparse-coding or deep-learning methods, can be generated on the same events for direct comparative evaluation.

Unless otherwise stated, we focus on one year of data over the entire 2012 calender year. Although SDO became operational in mid 2010, this is the earliest date currently available for the Trainable Module image parameter data. The single

[2]http://www.lmsal.com/isolsearch

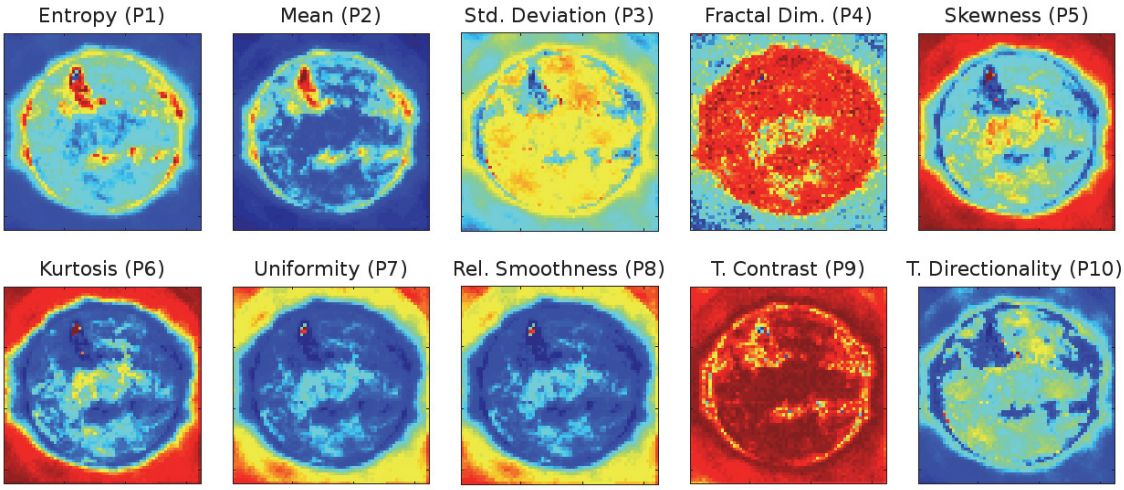


Figure 2: Image parameter heatmap plots for an AIA image, where each plot is normalized from 0.0 (blue) to 1.0 (red).

year dataset[3] contains approximately 38,000 event (data) instances, and while solar activity is not constant from year to year, we can safely estimate thousands of new events each year of additional data added in future works.

### 3.1 Region Labeling

The first step is to identify the spatiotemporal regions of interest (ROIs) for study. We focus on two specific event types of interest from the SDO FFT module reports: Active Region (AR) and Coronal Hole (CH). These were chosen in part because of similar reporting characteristics that make region identification easier. Over the year there is approximately 13,518 AR events and 10,780 CH events. Both event types are reported at approximately a four hour cadence, providing a snapshot of all visible event instances at each report time. We chose to use AR events as our base time for linking all data sources together. We first collect all unique AR report times, and then for each one, we find all CH events that are within $\pm$ 60 minutes. If no CH events exist, we skip the AR events to ensure each timestep has both event types. While it is not necessarily essential to have both types of events (labels) present at each timestep, it does mitigate the concern of unknowingly overlapping labels that would decrease the effectiveness of discrete label-based learning tasks.

Next, we calculate the average of all linked AR and CH events and create a new list of dataset events with their updated timestamps. These events include an ID link back to their raw reports for later lookup if necessary. They also contain the spatial information, which consists of a center point, bounding box, and polygonal outline (chain code). Since all AR and CH event instances contain chain codes, unlike many other event types, we use this spatial attribute for a more precise ROI. In all, we retain 2,116 unique event times and only discard 12 due to the above restrictions.

Additionally, we artificially create a third event type called Quiet Sun (QS), which represents areas of the solar disk where neither AR or CH events exist. These are generated by taking the bounding box of each AR event in our new list and randomly placing it somewhere else on the solar disk without overlapping any other AR or CH events at that time – another helpful reason to only retain times with both event types present. By using AR events as templates, we replicate the natural distribution of event sizes and frequencies, and create exactly the same amount of QS events (in this case 13,518). While we could use the chain code for these as well, it intuitively makes little difference and is computationally much more efficient to use bounding boxes.

### 3.2 Region Representation

After the list of events has been established, we now generate instances for the actual dataset. An example image with labeled ROIs can be seen in Figure 3, with the underlying image parameter $64 \times 64$ grid cells highlighted for each region. Notice the large variability possible between an event's bounding box and chain code, and the lack of chain codes for QS events. We first verify associated image parameter data is present within $\pm$ 30 minutes over all nine AIA channels, otherwise we mark the event skipped for a final tally of data coverage. Over the entire year, this only happens 137 times (roughly 6.5%), leaving us with a final count of 1,979 unique event times for the 37,816 event instances.

Each event instance contains a variable number of image cells. Each cell is represented by 10 parameters (see Fig. 2) that are extracted from each of the 9 AIA channels, creating in total 90 parameters (dimensions). As a preliminary work establishing the foundations and baseline of performance, we want to begin with a simple strategy that is easy to understand and not domain-specific. Therefore, we calculate a 7-statistic summary for each parameter over all cells in the ROI, ordered as: minimum, 1st quartile, median, 3rd quartile, maximum, average, and standard deviation, which we notate as the vector: {q0, q1, q2, q3, q4, avg, std} for shorthand abbreviation. Therefore, each event instance is repre-

[3] http://dmlab.cs.gsu.edu/solar/

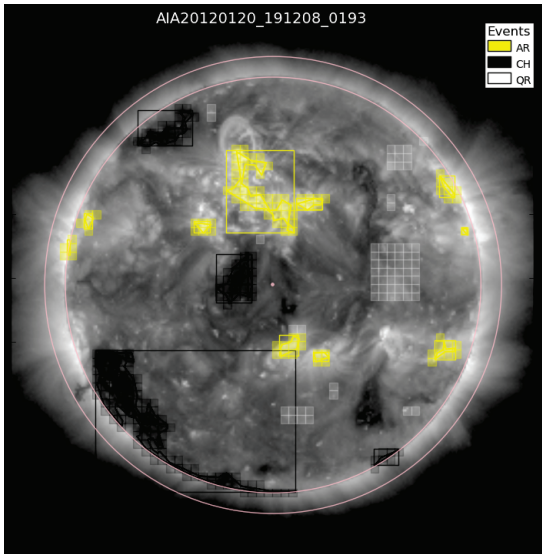


Figure 3: Example of cell-based event ROIs.

sented in a 630-dimensional space ($9 \times 10 \times 7$).

We note that the total ordering is mostly only important for interpreting results by IDs (1-630), which are a nested combination of: AIA channel (94,131,171,193,211,304,335,1600,1700), image parameter (1-10), and statistical values (1-7). So for example, ID 1 is the minimum of entropy in AIA 94Å written in shorthand as "0094-P01-q0", and so on.

In addition to these content-based features, we derive a number of metadata characteristics for each region to be used later for additional qualitative evaluation and possible filtering capabilities. While advanced use of this metadata is beyond the scope of this current work, we highlight them here as they are generated during the representation process for future use. Namely, we record the distance from solar disk center to event center, the radial cosine angle of the event based on disk center, the bounding box size, and the number of cells within the chain code. These attributes provide more specific information on the absolute location, size, and shape of each event instance.

# 4 Experimental Analysis

Now that we have generated a usable dataset, we follow several supervised experimental methods to evaluate our potential CBIR performance. We begin with feature selection to reduce dimensionality and then briefly look at classification accuracy to see how well we can learn the three event types (class labels) over various dimensionality. This is important because it provides us more insight into prominently selected features as well as more assurance in the feature selection process. Lastly, we perform k-nearest neighbor retrieval and evaluate precision for each individual event type. We also showcase several example query results and their associated meta data to further highlight the robustness (and non-trivial results) of our retrieval performance.

Unless otherwise stated, before the following experiments

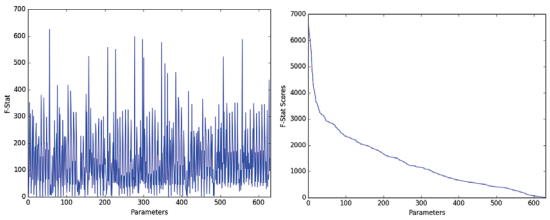


Figure 4: F-Statistic values for all 630 features, sorted by ID (left) and value (right).

are performed we balance the dataset through random under sampling. This leaves us with 10,780 instances for each event type, which is only a slight reduction to 32,340 total instances in the dataset.

## 4.1 Feature Selection

The first step is feature selection to reduce our dataspace from 630 original dimensions to something generally more manageable. Done properly, this can enhance classification and retrieval performance by better separating event types (data classes) and reducing computational burdens. It will also significantly speed up k-nearest neighbor retrieval through effective high-dimensional indexing techniques.

We use the F-statistic as a scoring (ranking) of feature importance towards class separation. Due to the greedy nature of this measure, we first randomize the balanced dataset. An example of F-stat values is shown in Figure 4, where the left plot is in original parameter order and the right plot is sorted by scores. We can see significant and repeated oscillations on the left, indicating certain image parameters (and/or derived 7-stat measures) are clearly more important than any specific AIA channel. On the right we see a common descending curve of F-stat values indicating the first features (approx. 40-60) are significantly more important than the rest in the selection process.

Table 1: Top 10 features based on 10-run F-Stat ranking.

| P-ID | P-Label | Rank | Score |
|---|---|---|---|
| 56 | 0094-P08-q0 | 3 | 6696.821 |
| 298 | 0211-P02-q4 | 14 | 6454.258 |
| 557 | 1600-P09-q4 | 24 | 6363.134 |
| 277 | 0193-P09-q4 | 28 | 6226.519 |
| 347 | 0211-P09-q4 | 40 | 6132.172 |
| 228 | 0193-P02-q4 | 49 | 6021.542 |
| 207 | 0171-P09-q4 | 60 | 5774.715 |
| 508 | 1600-P02-q4 | 70 | 5650.484 |
| 158 | 0171-P02-q4 | 79 | 5584.087 |
| 301 | 0211-P03-q0 | 87 | 5320.306 |

Preliminary empirical results showed that smaller datasets (e.g., 1 week and 1 month samples) have greater variability in resultant ranked features. Therefore, while one year provided more stable results, we also chose to perform ten separate runs and then sum their scores (and ranks) together for a final robust list. Table 1 lists the top 10 features sorted by their rank-sums. We note that the ranking values are 0-based,

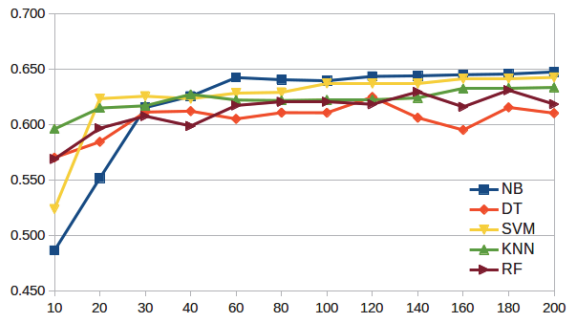


Figure 5: Classification accuracy over dimensionality.

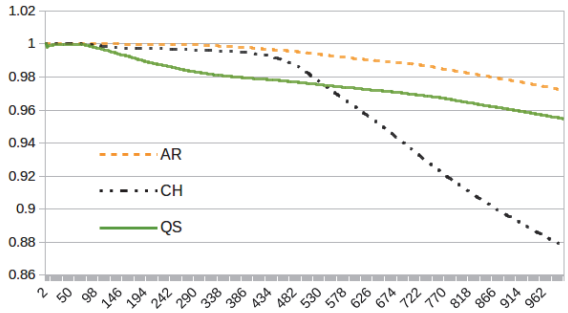


Figure 6: Precision results for KNN retrieval.

so for example, the first feature with a rank-sum of 3 indicates that over all ten runs it was at worst chosen 4th once (and 1st nine times), or at best 2nd three times and 1st the remaining seven times. Therefore we have high confidence this is indeed the most important feature in the dataset.

Interestingly, notice that 8 of the 10 are q4 (max) and coming from P02 (mean) and P08 (relative smoothness). This strongly indicates usefulness of these two image parameters, as well as the frequently used max-pooling strategy often used in windowing aggregations of more advanced works. We note that std (standard deviation) also appears very frequently in the top 50, and out of the entire top 100 features only five are not q0, q4, or std statistics.

## 4.2 Classification

Next we move to supervised classification using our ternary labeled dataset (AR, CH, QS). To gauge general classification performance, we evaluated: Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) methods. While rigorous tuning of these algorithms is beyond the scope and purpose here, we did empirically evaluate several critical parameters and show only the best here. This includes using Entropy criterion for trees with a max-depth of 10, RBF kernels for SVMs, and K=30 neighbors for KNN.

Using the list of ranked features discovered in the previous section, we evaluate successively larger dimensionality from 10 to 200, essentially appending more dimensions back on to the reduced space. In other words, the top 10 parameters shown above are always used. We perform standard 2/3 train, 1/3 test, but due to temporal dependence, rather than a randomized shuffle of data instances we split our training and testing sets temporally as 8 (and 4) months, respectively. This is an often overlooked, but critically important, aspect that can drastically affect certain learning algorithms. For example, if we have the same AR event represented four times, each only four hours apart, then there is a large possibility that the test set would contain one of these events that is likely very similar to the other events in the training set. During preliminary experiments we observed this situation could occur with this data and lead to inflated results.

Figure 5 displays the classification accuracy results for the five machine learning algorithms over dataset dimensionality. Note that all five algorithms appear quite stable between 60-65% accuracy, whereas a random guess would achieve 33% accuracy (3 classes). Some of the minor variances may be eliminated through more runs on larger datasets, but all algorithms appear to stabilize after about 40-60 dimensions, which is less than 10% of the original 630 dimensional space. Although additional dimensions may be slightly increasing accuracy of NB, SVM, and KNN, it is likely not worth the trade-off in computation for our purposes.

## 4.3 Retrieval

Lastly, we look at k-nearest neighbor (KNN) retrieval results to assess CBIR capabilities, and specifically Region-based Querying (RBQ) performance. We select 100 random data instances for each class and retrieve the nearest 1,000 neighbors using Euclidean distance on the top 60 dimensions. We then aggregate results over all queries for each class and calculate precision for each additional neighbor, which is defined as the ratio of true positives (same class label) over the total number of retrieved results. Because we do not extend retrieval results through the entire dataset, we discard the recall measure, which is essentially encompassed in precision performance on a limited KNN set.

In Figure 6, we present the precision results for each class averaged over 100 queries from K=1 to K=1,000 in 1-step increments. Remarkably, we find that AR queries maintain a perfect precision ratio upwards of K=150 and CH to about half of that at K=75. While perfect precision is not a necessity, it does indicate very good RBQ capability on what are likely highly distinguishable classes (event types). We note that QS begins to degrade in precision almost immediately at K=5, indicating less definable characteristics, which one would expect given the creation process.

Unlike the prior evaluations, RBQ retrieval uses the entire (un-balanced, non-separated) dataset. This provides a more real-world application, but might cause other issues. First, we note that there are many more events per class than our largest query, so exhausting the dataset will not be a factor (K=1,000 with over 10,000 events per class). Second, we could have temporal-dependence affects skew our results as suggested above. We briefly investigate this by examining the metadata features of top results for example queries that clearly indicates if we are returning "too-similar" results of spatiotemporal neighbors, which is a known issue for existing full-disk solar CBIR systems (Banda et al. 2013).

In Table 2 we present the metadata of the top 3 results for an example query from each event type. Empirically we observed similar results for many analyzed queries, but due to space we limit the results here to a small example. We note all the results match the same label as the query item. Notice the differences in query and result dates, indicating these results are not direct temporal neighbors. Also, we provide the center (X,Y) locations offset by disk center at (2048,2048), and the total grid cells contained in each region.

Here we can clearly see the results are not direct spatial neighbors and can vary in size. Therefore, we have strong confidence in our system returning truly similar (content-based feature) results independent of the actual spatiotemporal qualities of the regions. In the future, we can fuse similarity searches over these metadata qualities to satisfy more specific user-defined queries.

## 5 Conclusions and Future Work

This paper serves as the foundation for a new region-based retrieval framework for heliophysics imagery data. We present a preliminary approach for representing labeled regions of interest within images using general-purpose grid-based image parameters along with a supervised evaluation methodology for comparing alternative strategies in future works. Our results showcase exceptional dimensionality reduction capability as well as a promising baseline for retrieval performance across multiple event types. The dataset is available for easy comparative evaluation and benchmarking performance results against alternative works.

Several directions of future work are being pursued. Additional event types for supervised analyses, unlabeled patches for exhaustive image corpus coverage, and indexing for retrieval scalability. Additionally, more advanced region representations, such as sparse coding models and deep neural networks, are under active research and development.

## Acknowledgments

This project has been supported in part by funding from CISE, MPS and GEO Directorates under NSF award #1443061, and by funding from the LWS Program, under NASA award #NNX15AF39G.

Table 2: Metadata results from example queries.

| Q-K | ID | Date | X | Y | Cells |
|---|---|---|---|---|---|
| Q-AR | 23487 | 08-28 08:29 | -817 | 267 | 10 |
| K-1 | 19054 | 07-11 11:27 | 52 | -608 | 23 |
| K-2 | 33841 | 11-21 18:51 | -964 | 241 | 12 |
| K-3 | 22054 | 08-13 14:01 | -1017 | -536 | 12 |
| Q-CH | 30021 | 10-24 12:26 | 131 | -1134 | 27 |
| K-1 | 25901 | 09-21 13:58 | -1166 | 30 | 32 |
| K-2 | 29274 | 10-19 04:13 | -87 | -749 | 21 |
| K-3 | 30067 | 08-28 12:29 | 32 | -1129 | 25 |
| Q-QS | 36407 | 12-16 12:27 | 1315 | -860 | 35 |
| K-1 | 30073 | 10-24 20:26 | -795 | 1184 | 49 |
| K-2 | 37338 | 12-25 10:33 | 207 | 826 | 72 |
| K-3 | 19329 | 07-15 13:59 | 673 | 261 | 110 |

## References

Banda, J. M., and Angryk, R. A. 2010a. An experimental evaluation of popular image parameters for monochromatic solar image categorization. In *The 23rd Florida Artificial Intelligence Research Society Conf. (FLAIRS)*, 380–385.

Banda, J. M., and Angryk, R. A. 2010b. 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)*, 528–534.

Banda, J. M.; Schuh, M. A.; Wylie, T.; McInerney, P.; and Angryk, R. A. 2013. 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)*.

Hurlburt, N.; Cheung, M.; Schrijver, C.; Chang, L.; et al. 2012. Heliophysics event knowledgebase for the solar dynamics observatory (SDO) and beyond. In *The Solar Dynamics Observatory*. Springer. 67–78.

Kempton, D. J.; Schuh, M. A.; and Angryk, R. A. 2016. Towards feature selection for appearance models in solar event tracking. In *Artificial Intelligence and Soft Computing: 15th International Conference, ICAISC 2016*, 88–101. Springer International Publishing.

Lemen, J.; Title, A.; Akin, D.; Boerner, P.; et al. 2012. The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *Solar Physics* 275:17–40.

Martens, P. C. H.; Attrill, G. D. R.; Davey, A. R.; Engell, A.; Farid, S.; Grigis, P. C.; et al. 2011. Computer vision for the Solar Dynamics Observatory (SDO). *Solar Physics*.

Pesnell, W.; Thompson, B.; and Chamberlin, P. 2012. The solar dynamics observatory (SDO). *Solar Physics* 275:3–15.

Rimmele, T.; McMullin, J.; Warner, M.; Craig, S.; et al. 2015. Daniel K. Inouye Solar Telescope: Overview and Status. *IAU General Assembly* 22:2255176.

Schuh, M.; Angryk, R.; and Martens, P. 2015. Solar image parameter data from the SDO: Long-term curation and data mining. *Astronomy and Computing* 13:86–98.

Schuh, M. A.; Angryk, R. A.; and Martens, P. C. 2016. A large-scale dataset of solar event reports from automated feature recognition modules. *Journal of Space Weather and Space Climate* 6:A22.

Schuh, M.; Angryk, R.; Ganesan Pillai, K.; Banda, J.; and Martens, P. 2013. A large scale solar image dataset with labeled event regions. In *20th IEEE Int. Conf. on Image Processing (ICIP)*, 4349–4353.

Verbeeck, C.; Delouille, V.; Mampaey, B.; and De Visscher, R. 2014. The spoca-suite: Software for extraction, characterization, and tracking of active regions and coronal holes on euv images. *Astronomy & Astrophysics* 561:A29.

Withbroe, G. L. 2000. Living With a Star. In *AAS/Solar Physics Division Meeting #31*, volume 32 of *Bulletin of the American Astronomical Society*, 839.