When Too Similar is Bad: A Practical Example of the Solar Dynamics Observatory Content-Based Image-Retrieval System

Juan M. Banda, Michael A. Schuh, Tim Wylie, Patrick McInerney, and Rafal A. Angryk

Montana State University, Bozeman, MT 59717 USA
{juan.banda,michael.schuh,timothy.wylie,
patrik.mcinerney,angryk}@cs.montana.edu

Abstract. The measuring of interest and relevance have always been some of the main concerns when analyzing the results of a Content-Based Image-Retrieval (CBIR) system. In this work, we present a unique problem that the Solar Dynamics Observatory (SDO) CBIR system encounters: too many highly similar images. Producing over 70,000 images of the Sun per day, the problem of finding similar images is transformed into the problem of finding similar solar events based on image similarity. However, the most similar images of our dataset are temporal neighbors capturing the same event instance. Therefore a traditional CBIR system will return highly repetitive images rather than similar but distinct events. In this work we outline the problem in detail, present several approaches tested in order to solve this important image data mining and information retrieval issue.

1 Background and Motivation

Content-based Image-Retrieval (CBIR) systems are imperative in many research areas and industries where the amount of information to sort, search, and retrieve is greater than what is humanly feasible. CBIR systems are currently used across many diverse fields such as medical vision, video surveillance, law enforcement, facial recognition, tracking, and more [?,?,?,?].

How the CBIR systems are used and how they work also vary depending on the needs of the application. Some systems are designed to find identical visual characteristics, while others focus on finding structural similarity. Defining what is of interest is an important aspect of a CBIR system. This measure is application dependent and guides what techniques can be used and how the data is processed. CBIR systems are well-known with methods having been developed which index and define interest based on color features [?], shapes of certain objects [?], textures [?], etc.

The Solar Dynamic Observatory (SDO) mission was launched in 2011 and captures 70,000 images a day over 10 wavebands providing an unprecedented 1.5 TB a day of information about the Sun. The exponential growth of cross comparison between all images makes most standard methods of comparison infeasible.

Therefore we must address a more important issue that is often overlooked in most systems since none of them, to our knowledge, have to deal with the same volume of highly similar data.

When studying the Sun, solar physicists primarily study solar phenomena that we will refer to as solar events. The type of events that are of interest can vary greatly in size, duration, location, and in the colocation possibilities with other events. Thus, we have many different types of tasks to perform in this CBIR system that no other traditional system performs. Our initial attempts to tackle this ambitious problem consist of several methods:

- We will focus on traditional nearest-neighbor retrieval to fully exemplify our problem and decide on a few starting points for more potential experiments.
- We experiment on our CBIR system by using the labels to attempt to solve the temporal cadence issue. These labels all come from central location, Heliophysics Event Knowledgebase (HEK), where all the Feature Finding Team (FFT) modules report to.
- When searching for similar events, the most similar images are from the proceeding and succeeding timesteps which also contain the same event we are querying on. This makes finding patterns or similar singular events difficult. We can intuitively reduce this by increasing the cadence of the system. Unfortunately, this results in excluding many important short duration events such as solar flares as we will show in the experiments section.

In this paper we introduce the problem of finding similar events in a temporal dataset. We discuss several ways to approach the problem and show how effective they result on our test dataset, and we also lay out some possible future directions to improve CBIR systems that face these similar issues.

2 Experimental Setup

2.1 Image Parameters

As we have presented in our previous works, we use some of the more popular image parameters that are used in fields such as medical imaging, natural scene images, and traffic imaging [?,?,?]. We use a grid-based image segmentation with 4,096 cells per image shown to be the most effective on our solar images [?]. The ten image parameters that we have defined to be the most useful [?] are: Entropy, Mean, Standard Deviation, 3rd Moment (skewness), 4th Moment (kurtosis), Uniformity, Relative Smoothness (RS), Fractal Dimension, Tamura Direcctionality, and Tamura Contrast, more details on them can be found on [?].

2.2 Filtering Mask

To identify regions of interest for active solar events, we employ a simple intensity-based region growing technique [?]. Pixels of intensity greater than the 99.5 percentile for the image are selected as the 'seeds' of the regions [?]. The regions are then grown by iteratively adding any pixels of intensity above the 80th percentile of the image that are 8-way adjacent to the current region. Terminating once no more pixels can be added. Next we apply a radial filter to the image, eliminating all pixels that are not within a fixed distance (the Sun's radius) of the image center. Finally, we build the set of grid cells that contain one or more pixels of the remaining regions, resulting in the white images to the right in Figure 1.

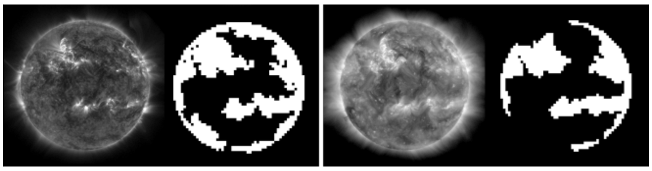

Fig. 1. Regions produced by the mask for wavelengths 171(left) and 193(right).

2.3 Similarity Measures

In order to determine if we can have more useful and interesting similarities between images/events, we utilize ten different distance metrics that will showcase different properties between our images/events. These metrics are addressed in detail in [?], and include: Euclidean, Std. Euclidean, City Block, Chebyshev, Cosine, Correlation, Spearman, and Factional Minkowski with $p = 0.5, 0.75, 0.90$.

2.4 The SDO Dataset

To create an SDO dataset we had to overcome one problem: finding annotated event data. Since asking experts to manually annotate 4k by 4k resolution images is unrealistic, we had to wait for several of the modules of the Feature Finding Team (FFT) of the SDO mission [?] to be fully running and reporting their results on their respective solar events they were designed to detect. This dataset consists of images from four different wavebands over a three-day period (from January 20, 2012 to January 23, 2012, subset of the one presented in [?]) where there was a representative amount of solar activity resulting in multiple occurrences for each type of event. We experiment with four different AIA wavebands (94, 131, 171, 193), and four types of labeled events: Active Region (AR), Coronal Hole (CH), Flare (FL), and Sigmoid (SG), with 292, 71, 161, and 95 event labels respectively. Each one of these events are reported by a module of the FFT that has been independently developed by a specialized team of solar physicists and computer scientists using image processing, statistical analysis, and data mining techniques [?].

The original version of this dataset can be found here [?]. In order to present our unique interestingness and relevance problem we will use four different versions of this dataset that are outlined in Table 1.

Label	Description	Images
Original	Four wavelengths, time cadence of 6 minutes	3,394
DS1	One image per labeled event	619
DS2	Original dataset with mask from section 2.2	3,394
DS3	One image per labeled event and mask from Section 2.2	619

Table 1. Original, DS2 and DS3 will be split by wavelength for our experiments.

2.5 Experiment Descriptions

Experiment 1 Using dataset Original, we calculate each images nearest neighbors for each different wavelength. We also performed the same calculation using all ten different distance metrics from section 2.3. Finally, we will plot all the images and their nearest neighbors sorted by their time stamp from left to right. With this experiment we show how the temporal aspect of our data is affecting the similarity problem, by returning as the closest neighbors all the closest temporal images.

Experiment 2 Using dataset DS1, we generate a similarity graph for each different event with all its possible nearest-neighbors from the same event type, generating a different graph for each distance metric. We then plot the events and the distances to others sorted by event time stamp from left to right. With this experiment we expect to see if the similarity effect of the temporal repetition of the images is reduced when we group sets of images in events, based on their time ranges. Ideally, we would see similarity plots that contain the most-similar events from the time range and not an ordered list of events by time stamp.

Experiment 3 Using dataset Original, we calculate each image nearest neighbors for each wavelength, but we add a time step component (sampling cadence). We also perform the same calculation using all ten different distance metrics. Finally, we plot all the images and their nearest neighbors sorted by their timestamp from left to right. By increasing the time cadence of sampled images from our dataset, we expect to lower the image repetition and have more interesting nearest neighbors than before. While seen as the most intuitive solution, this approach will introduce other problems.

Experiment 4 Using the same set-up as Experiments 1, 2, and 3, but with datasets DS2, DS3, and DS2, respectively, we again plot the images and their nearest neighbors sorted by time-stamp from left to right. In order to reduce the storage expense and remove uninteresting parts of the solar image in an automated way, we apply the mask described in Section 2.2 to get any performance gains from considerably reduced datasets that now only contains regions of interest.

3 Results and Analysis

This section contains the most interesting results for the previously outlined experiments. If you are interested in seeing all of the resulting plots, or in replicating the experiments, please visit the supplemental website [?].

3.1 Experiment 1

Testing every image in the dataset separated by wavelength will allow us to observe the temporal similarity relation between the images nearest neighbors. Our similarity (a.k.a nearest neighbor) matrix is plotted in a symmetrical way and all our distance values are scaled from 0 to 1, with 0 being closer and 1 being the farthest away. The colors of our plot range between dark blue to dark red and they represent the same 0 to 1 scale, respectively.

As we can see, the temporal similarity corresponds to the distance levels as they change from blue to red. The problem is that almost all behave in the same manner–it is blue when it is close temporally. From the range of the dark blue we can see the closest neighbors are the immediate temporal images. This

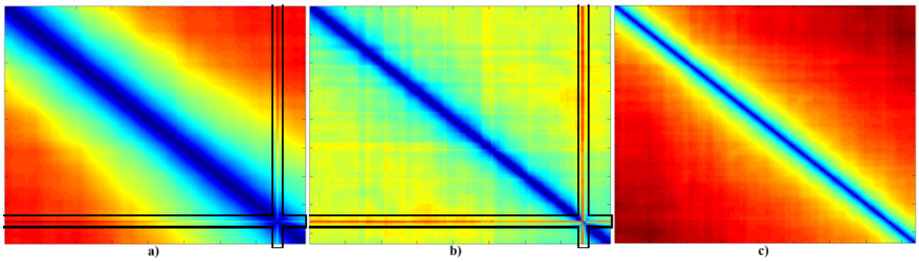

Fig. 2. Nearest neighbor plots for 131 a) Sperman distance and param. mean, b) Correlation distance param. tamura contrast 10. 171 c) City block distance and param. standard deviation.

behavior is consistent, except for the boxed region in Figure 2 a and b, which region/timeframe corresponds to a large solar flare. In this event the intensity of the solar values goes very high for a short period of time and drastically fluctuates between consecutive images and thus the red streaks inside the selected area. However, as the event ends, the behavior returns to normal again having the events be very similar in terms of distance with respect to time.

Keeping with the consistent behavior of shifting from blue to red as the time stamp increases, we show that almost any combination of image parameter and distance metric will be a victim of this similarity problem. The flare event outline is found on Figure 2 a, b, with only c (city block distance with param. standard deviation) eliminating its presence. This will quickly lead us to discarding the combination since we will lose an event we are trying to find.

3.2 Experiment 2

Taking one image per solar event, we can reduce our dataset from 3,394 to 619 images since each event has a duration window ranging from minutes to several hours depending on the FFT module and their event-specific reporting standards, more on this can be found in [?]. With such reduction, we expected to have the time repetition factor less apparent.

Figure 3 a) and b) show the problematic behavior with this approach. As expected, the diagonal is 0 or blue, and the spreading out pattern goes from blue to light green (around 0.29 according to the scale), which indicates that most active region events that are similar are consecutive (within their wavelength) in time which is not useful for researchers trying to find similar events at a different time. Note this plot features the events sorted by time-stamp, not similarity, and features two different wavelengths as seen from the two sections in a) and b).

It is worth mentioning that there are two events without any real or temporal nearest neighbors indicating a completely different event occurence. The flare outlined in Figure 2 is one example. From the labels provided by the FFT modules, we occasionally find inconsistencies like this. This furthers our emphasis of not relying only on labels for proper functionality of our CBIR system.

3.3 Experiment 3

The problem of temporally-dependent nearest neighbors is well showcased in Experiments 1 and 2 (and Figures 2 and 3). In our next experiment, we attempt

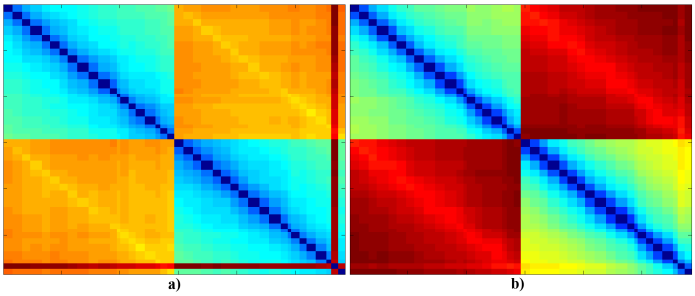

Fig. 3. Nearest neighbor plots for (by rows) AR a) Chebyshev distance and param. kurtosis, b) Euclidean distance param. standard deviation.

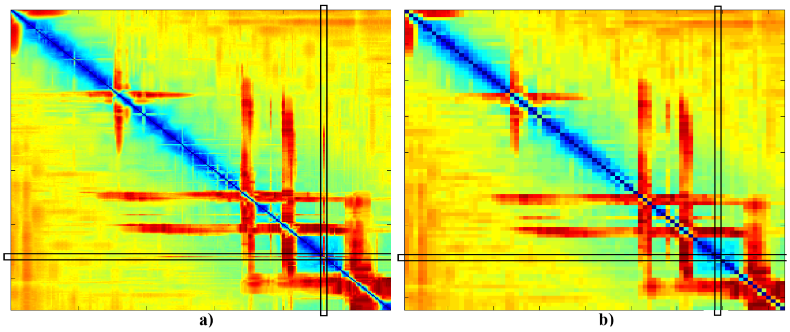

Fig. 4. Wavelength 94 nearest-neighbor plot for Chebyshev distance and entropy parameter with times step a) 18 minutes, and b) 60 minutes.

to solve this problem by increasing the time cadence of sampled images, from the original 6 minutes to 18 to 60 minutes respectively. The results are more promising than for the similarity plots, but introduce one very critical issue that we will outline in the following figures.

Increasing the cadence allows some events to completely disappear, causing our system to miss some potentially relevant results. As we can see in the event found in Figure 4 a). When the time cadence increases this event fully disappears on the 60 minute cadence (b). While increasing the time cadence is a naïve and intuitive solution to reduce temporal repetition in nearest neighbors in a traditional CBIR system, this may not be ideal or useful for our Solar CBIR system, since it causes problems for short lived solar events that occur rapidly.

3.4 Experiment 4

While the masked versions of the Original dataset provided nearly identical results when used in Experiment 1 and 3, the most interesting and revealing results came for the masked version of Experiment 2.

In Figure 5 we have a clear example of how the mask allows our similarity results to change. The sigmoid events reported on the 131 wavelength are all similar on the DS1 dataset, but after the mask is applied for DS3, we can now differentiate them effectively. This makes a strong case that when grouping by events, there is a benefit to using an image mask to determine regions of interest. Another interesting behavior is that we are now able to see the event similarities across wavelengths– something that before was unlikely. This can be seen when comparing the upper right quadrants of a) and b) in Figure 5.

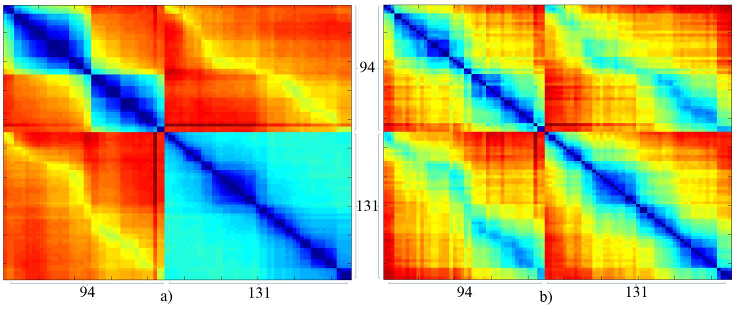

Fig. 5. SG event plot. DS1 dataset on the left, DS3 dataset on the right.

4 Conclusions

In this work we have outlined multiple intricacies of dealing with a dataset that has very similar images. Our problem also lies within the fact that our images are 4,096 by 4,096 pixels. When these two factors combine, we have determined that even when using different distance metrics, we will not be able to successfully analyze our image database without some kind of region of interest (ROI) approach that helps narrow down the query area and perform the image comparison at a lower level.

The results of Experiments 1, 2, and 3 indicate that the interestingness of our retrieval results cannot easily be solved by grouping images together based on event labels (Experiment 2) or by increasing the time cadence (Experiment 3). Since little literature exists that shows a CBIR system with this unique problem, it makes it an interesting data mining problem. There are few other known datasets that also have this problem, but we hypothesize that any video retrieval system would have a similar problem.

While the masked dataset DS2 did not provide any insightful results with Experiments 1 and 3, we did see an interesting development when used in conjunction with event based grouping (Figure 5 in Experiment 4). This leads us to believe that with the right combination of event type, wavelength, distance metric, and image parameter, we can still improve the differentiation of time-independent nearest neighbors.

We can conclude with the experiments performed and the analysis of the results, that with a hybrid approach combining time cadence reduction, interesting region mask, and the event grouping by waveband, distance metric, and parameter combination we would be able to see improvements in the returned nearest neighbors.

5 Future Work

We have started investigating practical and scalable solutions to region-based queries for our solar CBIR system. We expect the diversity of regions to diminish the problem of finding too similar results, but it will not be entirely eliminated. While we cannot use brute-force similarity-matching on dynamic regions for the full-scale system, it could provide us a benchmark of performance on a small sample dataset while working towards provably better solutions.

One possible direction for improvement is the incorporation of pre-defined region segmentation through clustering and classification on known data characteristics which could help reduce the search space of typical queries through pruning large quantities of unwanted regions. If a user is querying a ROI that resides in a bright region, we can eliminate all other regions while maintaining a high likelihood of returning similar region results with similar events.

More advanced solutions are still needed, and current directions of interest include using variations on recent visual bag-of-words approaches, or exploring hybrid indices that combine different data characteristics to achieve a singular comprehensive index. Thus, our existing full-image similarity-based indices must be extended to regions with spatial and temporal contexts.

References

1. R. Adams and L. Bischof. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(6):641–647, June 1994.
2. ADBIS. Adbis 2013 website. http://www.jmbanda.com/ADBIS2013, 2013.
3. Juan M. Banda and Rafal A. Angryk. On the effectiveness of fuzzy clustering as a data discretization technique for large-scale classification of solar images. In *IEEE Int. Conf. on Fuzzy Systems*, FUZZ-IEEE, pages 2019–2024, aug 2009.
4. Juan M. Banda and Rafal A. Angryk. Usage of dissimilarity measures and multidimensional scaling for large scale solar data analysis. In *Proc. of the 2010 Conf. on Intelligent Data Understanding*, CIDU'10, pages 189–203, Oct 2010.
5. Michael A. Schuh, Rafal A. Angryk, Karthik Ganesan Pillai, Juan M. Banda, and P. Martens. A Large-Scale Solar Image Dataset with Labeled Event Regions. In *20th IEEE Int. Conf. on Image Processing*, ICIP'13, to appear.
6. Juan M. Banda, Rafal A. Angryk, and Petrus C. Martens. On the surprisingly accurate transfer of image parameters between medical and solar images. In *18th IEEE Int. Conf. on Image Processing*, ICIP'11, pages 3669–3672, sept 2011.
7. A. Benkhalil, V. Zharkova, S. Zharkov, and S. Ipson. Active region detection and verification with the solar feature catalogue. *Solar Physics*, 235:87–106, 2006.
8. George Gagaudakis and Paul L. Rosin. Incorporating shape into histograms for cbir. *Pattern Recognition*, 35(1):81–91, 2002.
9. Feng Jing, Mingjing Li, and Lei Zhang. Learning in region-based image retrieval. In *Image and Video Retrieval*, volume 2728 of *Lecture Notes in Computer Science*, pages 199–204. Springer Berlin / Heidelberg, 2003.
10. S. Kulkarni and B. Verma. Fuzzy logic based texture queries for cbir. In *Proc. of the 5th Int. Conf. on Computational Intelligence and Multimedia Applications*, ICCIMA '03, pages 223–, Washington, DC, USA, 2003. IEEE Computer Society.
11. Zhang Lei, Lin Fuzong, and Zhang Bo. A CBIR method based on color-spatial feature. In *Proc. of the IEEE Region 10 Conf.*, volume 1 of *TENCON'99*, pages 166–169, 1999.
12. Hsin-Chih Lin, Chih-Yi Chiu, and Shi-Nine Yang. Linstar texture: a fuzzy logic cbir system for textures. In *Proc. of the 9th ACM Int. Conf. on Multimedia*, MULTIMEDIA '01, pages 499–501, New York, NY, USA, 2001. ACM.
13. P. C. H. Martens, G. D. R. Attrill, A. R. Davey, A. Engell, and et al. Computer vision for the Solar Dynamics Observatory (SDO). *Solar Physics*, 275:79–113, 2012.
14. Stefan Thumfart, Wolfgang Heidl, and et al. A quantitative evaluation of texture feature robustness and interpolation behaviour. In *Proc. of the 13th Int. Conf. on Computer Analysis of Images and Patterns*, CAIP'09, pages 1154–1161, 2009.