Unsupervised learning techniques for detection of regions of interest in Solar Images

Juan M. Banda and Rafal A. Angryk
Georgia State University
Department of Computer Science, P.O. Box 5060
Atlanta, GA 30302-5060
Email: jbanda@gsu.edu, angryk@cs.gsu.edu

***Abstract*—Identifying regions of interest (ROIs) in images is a very active research problem as it highly depends on the types and characteristics of images. In this paper we present a comparative evaluation of unsupervised learning methods, in particular clustering, to identify ROIs in solar images from the Solar Dynamics Observatory (SDO) mission. With the purpose of finding regions within the solar images that contain potential solar phenomena, this work focuses on describing an automated, non-supervised methodology that will allow us to reduce the image search space when trying to find similar solar phenomenon between multiple sets of images. By experimenting with multiple methods, we identify a successful approach to automatically detecting ROIs for a more refined and robust search in the SDO Content-Based Image-Retrieval (CBIR) system. We then present an extensive experimental evaluation to identify the best performing parameters for our methodology in terms of overlap with expert curated ROIs. Finally we present an exhaustive evaluation of the proposed approach in several image retrieval scenarios to demonstrate that the performance of the identified ROIs is very similar to that of ROIs identified by dedicated science modules of the SDO mission.**

I. Introduction

Searching through large-scale image repositories has been an open research topic over the years. With the current advances in deep learning making the most promising strides in the field [1], the problem of finding regions of interest (ROI) to minimize the search space becomes a more important task. In our particular application, the Solar Dynamics Observatory (SDO) mission, where the current solar image dataset spans over 45 million images since the mission launched in 2010, solar physicist sit on a treasure trove of information that is underutilized. Researchers have advocated and made great advances in the field [2], by developing scalable big data content-based image retrieval (CBIR) systems [3] using descriptor signatures [4]. This however had limited success for ROIs. The growing amount of image data and interest of the solar physics community, makes exhaustive full-image searches very expensive or unfeasible in a reasonable amount of time while full-image search has shown to be not as useful for solar physicist as region-based search can be.

Due to the surprising similarity of solar images with radiography medical imaging [5], we analyzed how medical imaging researchers have been extracting ROIs from their images [6], [7] and decided to use unsupervised learning methods (clustering) for our purposes. Experimenting with clustering algorithms such as K-Means [8], K-Medoids [9] and Expectation-Maximization (EM) [10], we introduce three evaluation metrics to determine the amount of overlap between the cluster-proposed ROIs and a pre-defined large set of ground truth labels extracted from the Heliophysics Events Knowledgebase (HEK) that correspond to four types of different solar events. We utilized the pre-defined ground truth labels in order to determine the cluster centers found in the data that overlap with the labels and used them as a newly generated ROI. This approach turns the unsupervised learning methods into semi-supervised learning due to the use of pre-existing labels after clustering to select the cluster centers of interest.

With an efficient mechanism in place [4] our work attempts to provide a more robust mechanism to detect ROIs real-time and reduce the querying space by a considerable amount (between 20% and 30% of an image) while still maintaining acceptable retrieval precision averages of around 75% for four different solar events.

The overall organization of this paper is as follows: after some background information (section II), we explain our evaluation methodology (section III) and show our experimental results (sections IV, V and VI). We discuss the experimental results in sections VI. Lastly we provide our conclusions (section VIII) and conclude the paper with an outline of our future work (section IX).

II. Background Information

The content-based image retrieval (CBIR) systems have been around since the 1995 when large image repositories started to be used for basic similarity searches that involved the querying actual contents of an image rather than a user-defined set of labels and meta-data. These early systems are extensively described in [11]–[13], however, CBIR systems in the field of Solar Physics did not exist until [5], [14] were developed for the SDO mission and publicly released in 2011. One characteristic that early CBIR systems have in common is that not until SIMPLIcity [15], Netra [16] and Walrus [17] were developed, they all only offered full-image querying. As the need for querying selected regions of images grew, researchers focused on adapting existing algorithms from CBIR into the Region-Based realm without addressing the scalability issue [18]–[21]. This can be explained by the fact that the first image repositories where not as large as SDO. The field of region-based image retrieval is efficiently summarized in [22], with the best performing approaches [23] relying on features involving dominant color, thus not transferring to the solar image domain. There are very few approaches that combine both ROI detection and clustering presented in [24], [25],

however, most approaches depend on ROIs with well visible objects that clearly contrast from the background [26], [27].

A. Dataset

We selected a subset of solar images compiled by Schuh et al. [28] from the Atmospheric Imaging Assembly (AIA) module of the SDO mission. This dataset spans a six-month period of data, and contains 17,785 labels (ROIs) from two separate wavelengths (193Å and 131Å), extracted from HEK. Details of the label counts are found in Table I.

TABLE I. DATASET DESCRIPTION

Event Type	Label	Wavelength	Total
Active Region	AR	193Å	7,108
Coronal Hole	CH	193Å	4,702
Flare	FL	131Å	4,316
Sigmoid	SG	131Å	1,659

This dataset has been analyzed in [4], allowing us to used the author's baseline retrieval results as a basis for our comparisons. Similar to Banda and Angryk [29], we will only use the Maximal Bounding Boxes (MBR) labels as using more refined labels (chain codes) has been shown to add extra computation expense with very minimal performance gain.

B. Image Parameter Extraction

Images are broken down in 64-by-64 individual cells of 64-by-64 pixels each, as shown in Figure1. Ten numerical image parameters: entropy, fractal dimension, mean intensity, the third and fourth moments, relative smoothness, the standard deviation of intensity, Tamura contrast, Tamura directionality and uniformity, are extracted from each cell. These parameters have been previously validated for solar images in [3], [30], [31].

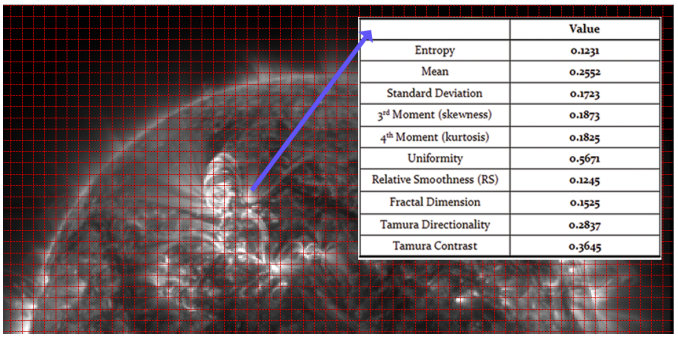

Fig. 1. Solar image with segments outlined and image parameter values extracted from an individual cell.

C. Baseline Image Retrieval System

In order to asses the value of our selected ROIs, we compare them with a Euclidean distance-based system introduced in [4] for region-based retrieval of solar images using signature descriptors. This system retrieves similar images based on ROIs with a base-line retrieval accuracy of: AR: 72%, CH: 85%, FL: 65% and SG: 63%, with an average of 71%, for the all the original event ground truth labels introduced by Schuh et al. [28].

D. Clustering Algorithms

In this analysis we tested three popular clustering algorithms, two exclusive clustering algorithms (K-Means, K-Medoids) and one distribution based algorithm (Expectation Maximization). We used the Fuzzy Clustering and Data Analysis Toolbox [32] for both K-means and K-medoids via Matlab, and Environment for Developing KDD-Applications Supported by Index-Structure (ELKI) [33] for Expectation Maximization clustering. One common requirement for the algorithms we selected is that they all take as an input a target number of clusters; this will allow us to compare them against each other more fairly.

K-Means [8]. This algorithm aims to partition n observations into k clusters (k is user-specified). For each iteration, the cluster with the nearest mean will be assigned to each individual observation, serving as a prospective cluster until convergence or a stop-criteria is reached. Defined as an NP-hard, problem there are many efficient heuristic algorithms that converge quickly to a local optimum. Each solution is not guaranteed to be unique. This algorithm will return k-cluster centers that will be used to determine the ROIs.

K-Medoids [9]. Similarly to K-Means, it attempts to minimize the distance between data points labeled to be in a cluster and a point designated as the center of that cluster. The main difference of this algorithm is that it selects actual data points as cluster centers (medoids) and thanks to that, it can use arbitrary distance metrics between data points (rather than Euclidean distance commonly used on K-Means).

EM [10]. This algorithm finds clusters by determining a mixture of Gaussians which fit the provided dataset. Each Gaussian has an associated co-variance matrix and mean. We used a variance scalar since we used spherical Gaussians for our experiments. The prior probability for each Gaussian is a fraction of points in the cluster defined by that Gaussian. These parameters can be initialized by randomly selecting means of the Gaussians with the algorithm converging locally to an optimal solution by iteratively updating values for means and variances.

E. Hardware Utilized

All of our experiments have been tested on separate virtual machines (VM), all running Ubuntu Linux 12.04 with 30 GB of RAM and using 6 cores of an AMD FX-8150 Black Edition processor. We provided the same hardware conditions for each system to run to be able to deliver 1-1 comparisons in terms of performance.

III. Evaluation Methodology

Since we are evaluating clustering effectiveness in terms of finding overlapping ROIs with our ground truth labels, and the image retrieval precision using ROIs, we outline two different evaluation methodologies below.

A. Clustering performance evaluation

Using the image parameters indicated in section II-B, we will now attempt to determine how many clusters and which clustering algorithm is the most efficient at finding ROIs. To do this we will measure the level of overlap of image cells tagged

with a certain cluster center against the image cells covered by the ground truth labels from [28]. This evaluation methodology allows us to identify which algorithm and cluster center(s) provide image cells that present possible ROIs for images that are unlabeled in the future. Figure 2 shows a ground truth label in (a) and the assigned cluster center for the same cell in (b). This is where the overlap counts will be calculated using the formula specified by Eq. 1 for each ROI. After calculating Eq. 1 for all cluster centers we determine the one with the highest value (i.e overlap) and use it to calculate Eq. 2. We lastly calculate the overlap rate between the correctly labeled image cells and the incorrectly labeled ones using Eq. 3.

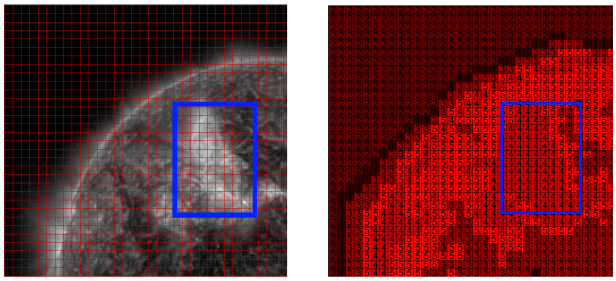

Fig. 2. Image used for retrieval evaluation: a) Grided image example with a ground truth label (in blue), b) Cluster center values are plotted inside each grid cell with a ground truth label (in blue) overlay.

$$correct\_overlap = \frac{\#\ cells\ in\ ground\ truth\ label}{\#\ cells\ with\ matching\ cluster\ center} \quad (1)$$

$$error\_overlap = \frac{\#\ cells\ outside\ of\ ground\ truth\ label}{\#\ cells\ from\ top\ cluster\ center} \quad (2)$$

$$rate = \frac{correct\_overlap}{error\_overlap} \quad (3)$$

B. Image retrieval evaluation

In order to evaluate the image retrieval performance of the potential ROIs discovered by our clustering algorithms, we will calculate descriptor signatures (calculated as outlined on Algorithm 1) for those sections. This approach has been validated in the past by Banda et al. [34] showing solid performance for the same solar image data we are using in this study. Each descriptor signature is represented by a histogram-like structure with ten bins, one for each extracted image parameter based on the average value of the image cells contained within said event boundary. The calculation process is outlined by Algorithm 1 (as previously published in [34]). This process reduces a complete ROI label into a robust 10-dimensional object making the retrieval calculation quite efficient, precision is then calculated via Algorithm 2.

IV. Experimental Results: Clustering

In order to determine the proper number of clusters needed to produce relevant ROIs we tested on *k* values between 2 and 16, in increments of 2. As researchers have shown in the past [31], it is more effective to evaluate clustering on individual image parameters rather than on the combination of them.

Algorithm 1 *Steps for calculating descriptor signatures [34]*

1: Calculate the maximum $Max(P_i)$ and minimum value $Min(P_i)$ of each of the 10 image parameters, for all cells in the dataset. Where P_i is the i-th image parameter value.
2: Match the boundary outline (blue MBR in Fig.2 of each event to the corresponding image cells. For each cell, find the parameter values.
3: Min-Max normalize each parameter value using: $P_i = \frac{P_i - Min(P_i)}{Max(P_i) - Min(P_i)}$
4: Take the average of each parameter P_i and use it in the bin value, in a histogram representing a given event.

Algorithm 2 *Retrieval precision calculation [4]*

1: Let E_i be the number of instances for the ith event type (eg. for Active regions we have 7,108 as specified in Table I).
2: Calculate the top E_i nearest-neighbors of each event.
3: Determine how many of them are of the ith event type, called true positives (TP).
4: Divide TP over E_i and multiply by 100. This results in the final accuracy percentage for that particular event.

Another parameter in our evaluation is that we will individually compare each solar event from the labeled dataset. This will allow us to capture the independent gray scale intensity of each event.

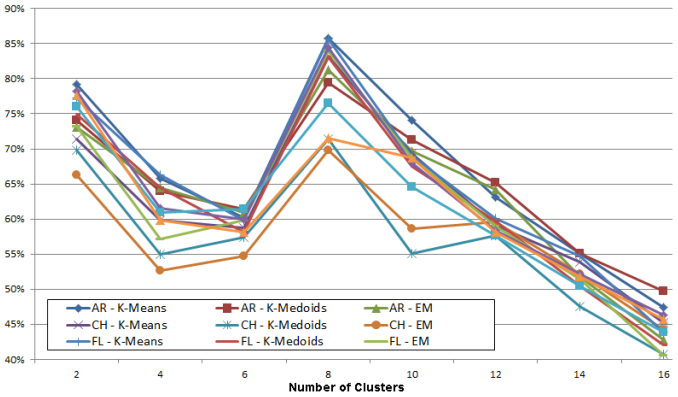

Fig. 3. Percentage of correct clustering ROI overlap versus ground truth image labels.

Figure 3 shows a plot of the calculated values for Eq. 1. We averaged the resulting value for every label per solar event from Table I. This figure indicates the number of correct cells found by the best performing cluster. By correct cell overlap we are referring to an overlapping cell that is found both in the ground truth label and the proposed ROI. In order to objectively determine which cluster selection is more efficient, we also look at the total number of cells that the same cluster labeled incorrectly as calculated by Eq. 2.

In Figure 4 we can see the amount of miss-categorized (error) cells based on the identified clusters to determine our potential ROIs. This number by itself is not very relevant as an unsupervised method will produce a considerable amount of false positives. However, this number is used to determine the

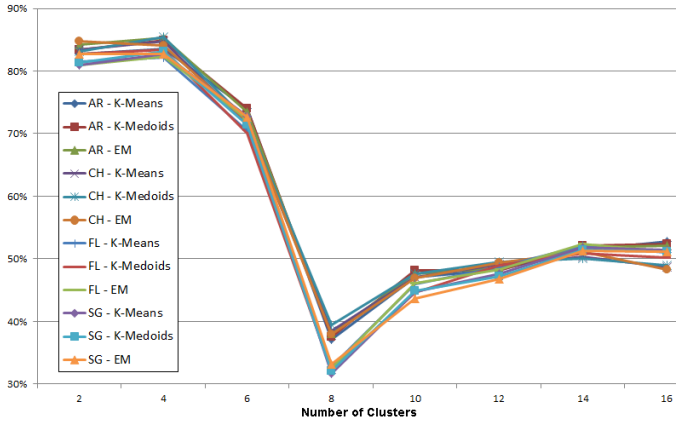

Fig. 4. Percentage error of clustering ROI overlap versus ground truth image labels.

rate of how many improperly labels cells are found versus how many properly labels cells did the clustering algorithm. This rate has been calculated on Table II. In order to determine

TABLE II. RATE (EQ.3) OF CORRECT VS. ERROR OVERLAP CALCULATIONS - THE BEST RESULTS ARE IN ITALICS

# OF CLUSTERS	4	6	8	10	12	14	16
AR - K-Means	0.79	0.81	*2.31*	1.57	1.31	1.07	0.90
CH - K-Means	0.71	0.82	*2.16*	1.47	1.20	1.07	0.93
FL - K-Means	0.81	0.85	*2.61*	1.50	1.23	1.05	0.86
SG - K-Means	0.74	0.82	*2.66*	1.52	1.23	1.01	0.90
Average	**0.76**	**0.83**	***2.44***	**1.52**	**1.24**	**1.05**	**0.90**
AR - K-Medoids	0.75	0.83	2.11	1.48	1.36	1.06	0.95
CH - K-Medoids	0.64	0.80	1.81	1.16	1.16	0.95	0.83
FL - K-Medoids	0.77	0.83	2.59	1.51	1.22	0.99	0.84
SG - K-Medoids	0.73	0.86	2.38	1.44	1.22	0.98	0.86
Average	**0.73**	**0.83**	**2.22**	**1.40**	**1.24**	**0.99**	**0.87**
AR - EM	0.76	0.83	2.14	1.46	1.32	1.00	0.82
CH - EM	0.63	0.76	1.84	1.25	1.21	1.02	0.91
FL - EM	0.69	0.83	2.58	1.49	1.22	0.98	0.79
SG - EM	0.72	0.80	2.16	1.57	1.24	1.01	0.89
Average	**0.70**	**0.80**	**2.18**	**1.44**	**1.25**	**1.00**	**0.86**

which cluster algorithm and number of clusters we will use for our retrieval experiments, on Table II we identify which combination of algorithm and number of clusters has the highest Eq. 1 score.

V. EXPERIMENTAL RESULTS: IMAGE RETRIEVAL

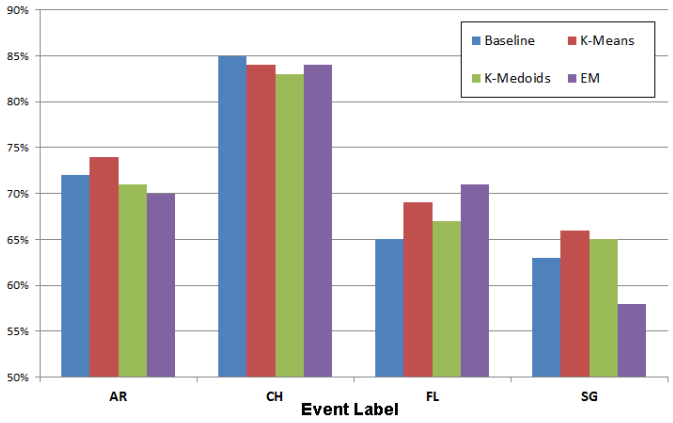

Fig. 5. Experimental image retrieval precision percentages.

While it is very clear that 8 clusters has the best rate of overlap vs. error, the difference between clustering algorithms is not that clear. As a way to determine ROIs in solar images, we proceeded to experiment by creating image descriptor signatures and test retrieval precision for them using the system developed by Banda et al. in [4]. The purpose of this evaluation is to verify if the retrieval of solar images using the cluster-produced ROIs is comparable to retrieval on the ground truth labels. In the process we also determined that K-means provided the best overall retrieval precision results (see Figure 5), with Table III showing the retrieval precision changes between the baseline system and the ground truth labels vs. the cluster-generated ROIs for all algorithms.

TABLE III. RETRIEVAL PRECISION IMPROVEMENT RATES

Algorithm	K-Means	K-Medoids	EM
AR	2%	-1%	-2%
CH	-1%	-2%	-1%
FL	4%	2%	6%
SG	3%	2%	-5%
Average	**2%**	**0%**	**-1%**

VI. EXPERIMENTAL RESULTS: TWO-CLUSTER ROI GENERATION

After visual inspection of multiple solar images plotted with their ground truth labels and compared side by side the cluster center assignments, we determined that we might gain some improvements by using more than one cluster for the ROIs generation. From Figure 6 we can clearly see in b) that two clusters are the ones that compose the majority of the ROI represented by the ground truth label. In the following experiments we selected the best performing algorithm (K-Means) and the best performing number of clusters (8) (as shown in Table III), and combined the top 2 clusters to generate the potential ROIs.

Table IV show the results of these changes. A considerable decrease of the correct overall vs. the error overlap rate (Eq. 3) is expected due to the fact that we will be capturing more cells outside the ground truth label of interest with two clusters, but this is justified with an improvement over the correct overlap calculations.

TABLE IV. RATE OF CORRECT VS. ERROR OVERLAP CALCULATIONS

Event	AR	CH	FL	SG
1-Cluster - Correct	85.78%	83.16%	85.67%	84.49%
1-Cluster - Error	37.09%	38.51%	32.78%	31.74%
Rate	**2.31**	**2.16**	**2.61**	**2.66**
2-Cluster - Correct	89.15%	85.17%	86.17%	84.69%
2-Cluster - Error	39.15%	38.74%	34.16%	31.78%
Rate	**2.28**	**2.20**	**2.52**	**2.66**

After finding an improvement of the number of labeled cells, we now have a plausible reason to explore image retrieval experiments on the two-cluster generated ROIs. Figure 7 shows the original baseline retrieval precision and compares it with the 1-cluster and the 2-cluster generated ROIs.

VII. DISCUSSION

The relevance of our experimental section lies in the fact that we want to find an automated method for ROI identification to stop depending on expert-curated ground truth labels.

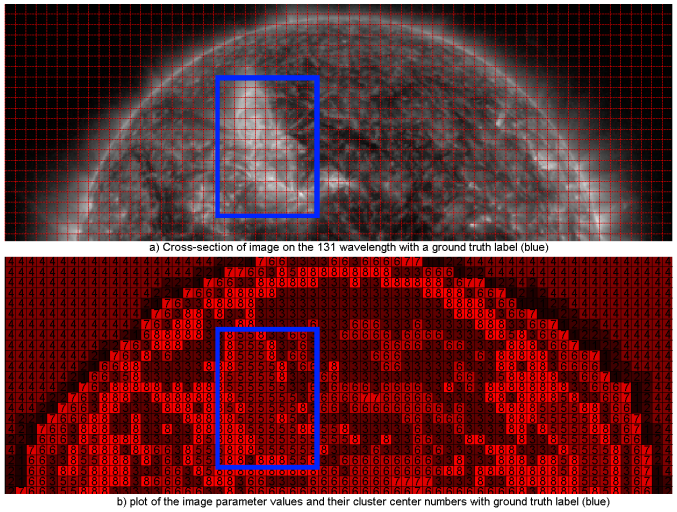

Fig. 6. Clustering visualization showing the which cluster centers overlap within a ground truth label (blue MBR).

The actual retrieval precision percentages and any inherent improvements in them are secondary in this works context. As we show in Tables II and III, our clustering approach does a fine job in covering the ground truth labels using 8 clusters consistently for all four types of solar events and for both the correct overlap evaluation (Figure. 3). When analyzed in parallel with the error overlap evaluation (Figure. 4) we get a rate of 2.44 on average (Tab. II) nearly one full point more than with any other number of clusters, further justifying our selection. This indicates that when we cluster our data using a *k*-parameter of 8 we have proportionally twice as many cells labeled properly than incorrectly, meaning that the selected cluster centers are very effective in determining the neighborhood of image parameters contained within the ground truth label. In terms of the best performing clustering

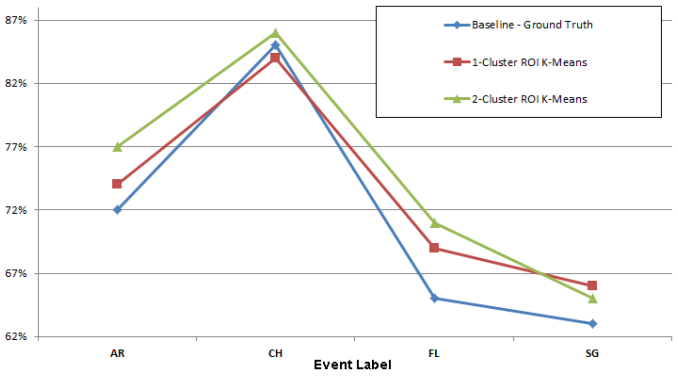

Fig. 7. Retrieval precision percentages comparison between baseline system with 1-cluster and 2-cluster results.

algorithm, we have very similar overlap performance for all three of the algorithms tested as shown in Figure. 5). Table III further clarifies the details. It is worth mentioning that while we only achieve a 2% improvement by using K-means, it is also the algorithm that took the least amount of execution time.

Moreover, the point of this work is not to be able to reproduce the known ROI labels, but rather develop the mechanism for automated ROI identification in images that have never been analyzed. These results suggest that we may be able to do this automatically for all remaining SDO images that never had ROIs identified and with comparable accuracy to the original event-recognition modules that produced the analyzed ground truth labels.

With the help of visualizations introduced in [35] we found that more than one cluster seems to provide better ROI overlap when compared against the ground truth labels (Figure. 6). This led to more complex experimentation scenarios, which demonstrated that by using the top two clusters we get better correct overlap percentages using only the top cluster in exchange for a very small reduction in the correct versus error overlap rate calculation, as seen in Table IV and Figure 7. When applied in a retrieval setting, this results in minimal retrieval precision improvements, but more importantly it clearly makes the point that we can automatically determine ROIs without the need for human or computer generated expert labels (Figure 7).

VIII. Conclusions

We have demonstrated that by using clustering algorithms we can identify the number of clusters needed to automatically generate ROIs for image retrieval. By putting the emphasis in the rate of correct versus error overlap between clustered ROIs and ground truth labels, we are able to achieve a balanced compromise and automatically generate ROIs with very little computational expense. Due to the nature of the SDO solar images with their grayscale range and fast changing solar events, our approach was able to achieve good results using a well-automated approach. At this point we do not make any claims that the clustering is actually finding solar events, but rather only interesting regions in an image that coincide with previously generated solar event labels treated as ground truth. By having an error overlap of 35% in the best case, we are still finding on average a considerable amount of new and unlabeled sections of the image that could be of interest for researchers while reducing the search pace to an average of 20% to 30% for each full image Here is where the premise of using a region-based image retrieval system comes into place, as it will be able to greatly take advantage of reducing the search space to only certain parts of the image rather than doing an exhaustive search.

The work presented here will greatly benefit the retrieval system developed in [3] and adapted for ROIs in [4], since once trained properly, it will allow the system to query massive image repositories by only having a limited amount of pre-existing image labels and without performing exhaustive searches on full images. For best performance our approach needs a considerable sample of ground truth labels to properly determine the top cluster centers assuming the image quality is always the same.

The main drawback of our approach is that it is sensitive to every different kind of solar event we tested, as the cluster center values for events such as Flares are very different than the ones for Filaments. However, with the proper filters implemented in the retrieval system we hope to be able to have the correct cluster centers calculated apriori and use them according to the types of queries submitted by researchers. In the worst case, the system will try all of them and try

to computationally determine the closest matches based on signature descriptor similarity, in a completely unsupervised manner.

IX. Future Work

Having developed our methodology and validated it for solar images, we will test it for medical images, specially radiographs. Researchers have shown that solar images and medical radiographs behave very similarly [5] for the group of image parameters we have used in this analysis. Another avenue to explore is the usage of the same methodology for spatio-temporal tracking of solar events, since we can adapt our method to keep track of the position of the potential ROIs and compare the results against artificially generated datasets, as generated by [36].

We first plan to implement our unsupervised ROI finding methodology as a module for [3] and deploy it live on [37] in order to determine if the results for images with no ground truth are satisfactory for solar physicists. While we have demonstrated that our approach has significant alignment with existing labeled images, we need to verify the methodology in an ad-hoc scenario. We plan on designing multiple usability tests to quantify the level of usefulness of our approach.

References

[1] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 157–166.

[2] J. M. Banda, M. A. Schuh, R. A. Angryk, K.-G. Pillai, and P. McInerney, "Big data new frontiers: Mining, search and management of massive repositories of solar image data and solar events," in *New Trends in Databases and Information Systems*, ser. Advances in Intelligent Systems and Computing, B. Catania, T. Cerquitelli, S. Chiusano, G. Guerrini, M. Kampf, A. Kemper, B. Novikov, T. Palpanas, J. Pokorn, and A. Vakali, Eds., vol. 241. Springer International Publishing, 2014, pp. 151–158.

[3] J. M. Banda, R. A. Angryk, and P. C. Martens, "Imagefarmer: Introducing a data mining framework for the creation of large-scale content-based image retrieval systems," *International Journal of Computer Applications*, vol. 79, no. 13, pp. 8–13, October 2013.

[4] J. Banda and R. Angryk, "Large-scale region-based multimedia retrieval for solar images," in *Artificial Intelligence and Soft Computing*, ser. Lecture Notes in Computer Science, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, and J. Zurada, Eds., vol. 8467. Springer International Publishing, 2014, pp. 649–661.

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

[6] R. A. Poldrack, "Region of interest analysis for fmri," *Social Cognitive and Affective Neuroscience*, vol. 2, no. 1, pp. 67–70, 2007.

[7] A. Elmoufidi, K. El Fahssi, S. Jai-Andaloussi, N. Madrane, and A. Sekkaki, "Detection of regions of interest in mammograms by using local binary pattern, dynamic k-means algorithm and gray level co-occurrence matrix," in *Next Generation Networks and Services (NGNS), 2014 Fifth International Conference on*, May 2014, pp. 118–123.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, p. 281297. [Online]. Available: http://projecteuclid.org/euclid.bsmsp/1200512992

[9] L. Kaufman and P. Rousseeuw, "Clustering by means of medoids," *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pp. North–Holland, 1987.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. pp. 1–38, 1977.

[11] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996.

[12] V. Ogle and M. Stonebraker, "Chabot: retrieval from a relational database of images," *Computer*, vol. 28, no. 9, pp. 40–48, Sep 1995.

[13] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the qbic system," *Computer*, vol. 28, no. 9, pp. 23–32, Sep 1995.

[14] J. Banda, R. Angryk, and P. Martens, "On dimensionality reduction for indexing and retrieval of large-scale solar image data," *Solar Physics*, vol. 283, no. 1, pp. 113–141, 2013.

[15] J. Wang, J. Li, and G. Wiederhold, "Simplicity: semantics-sensitive integrated matching for picture libraries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 9, pp. 947–963, Sep 2001.

[16] A. Natsev, R. Rastogi, and K. Shim, "Walrus: a similarity retrieval algorithm for image databases," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 3, pp. 301–316, Mar 2004.

[17] F. Jing, M. Li, L. Zhang, H.-J. Zhang, and B. Zhang, "Learning in region-based image retrieval," in *Image and Video Retrieval*, ser. Lecture Notes in Computer Science, E. Bakker, M. Lew, T. Huang, N. Sebe, and X. Zhou, Eds., vol. 2728. Springer Berlin Heidelberg, 2003, pp. 206–215.

[18] W.-Y. Ma and B. Manjunath, "Netra: A toolbox for navigating large image databases," *Multimedia Systems*, vol. 7, no. 3, pp. 184–198, 1999.

[19] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "An efficient and effective region-based image retrieval framework," *Image Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 699–709, May 2004.

[20] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "Region-based relevance feedback in image retrieval," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, vol. 4, 2002, pp. IV–145–IV–148 vol.4.

[21] J. Li, J. Z. Wang, and G. Wiederhold, "Irm: Integrated region matching for image retrieval," in *Proceedings of the Eighth ACM International Conference on Multimedia*, ser. MULTIMEDIA 00. New York, NY, USA: ACM, 2000, pp. 147–156.

[22] W. Huang, Y. Gao, and K. Chan, "A review of region-based image retrieval," *Journal of Signal Processing Systems*, vol. 59, no. 2, pp. 143–161, 2010.

[23] N. Shrivastava and V. Tyagi, "Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching," *Information Sciences*, vol. 259, pp. 212 – 224, 2014.

[24] S. Zakariya, R. Ali, and N. Ahmad, "Combining visual features of an image at different precision value of unsupervised content based image retrieval," in *Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on*, Dec 2010, pp. 1–4.

[25] A. Amory, R. Sammouda, H. Mathkour, and R. Jomaa, "A content based image retrieval using k-means algorithm," in *Digital Information Management (ICDIM), 2012 Seventh International Conference on*, Aug 2012, pp. 221–225.

[26] J. Amores, N. Sebe, and P. Radeva, "Context-based object-class recognition and retrieval by generalized correlograms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 10, pp. 1818–1833, Oct 2007.

[27] M. A. Helala, M. M. Selim, and H. H. Zayed, "A content based image retrieval approach based on principal regions detection," *International Journal of Computer Science Issues*, vol. 9, no. 4, pp. 204–213, 2012.

[28] M. Schuh, R. Angryk, K. Pillai, J. M. Banda., and P. Martens, "A large-scale solar image dataset with labeled event regions," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 4349–4353.

[29] J. M. Banda and R. A. Angryk, "Scalable solar image retrieval with lucene," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 11–17.

[30] J. M. Banda and R. A. Angryk, "An experimental evaluation of popular image parameters for monochromatic solar image categorization." *FLAIRS Conference*, 2010.

[31] J. M. Banda and R. A. Angryk, "On the effectiveness of fuzzy clustering as a data discretization technique for large-scale classification of solar images," *IEEE International Conference on Fuzzy Systems - 2009. FUZZ-IEEE 2009*, pp. 2019–2024, 2009.

[32] B. Balasko, J. Abnyi, and B. Feil. (2005, July) Fuzzy clustering and data analysis toolbox for use with matlab, university of veszprem. [Online]. Available: http://www.fmt.vein.hu/softcomp/fclusttoolbox/

[33] E. Achtert, H. Kriegel, and A. Zimek, "Elki: a software system for evaluation of subspace clustering algorithms," in *Scientific and Statistical Database Management, 20th International Conference, SSDBM 2008, Hong Kong, China, July 9-11, 2008, Proceedings*, 2008, pp. 580–585.

[34] J. M. Banda, C. Liu, and R. A. Angryk, "Region-based querying of solar data using descriptor signatures," in *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, ser. ICDMW '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1–7.

[35] M. A. Schuh, J. M. Banda, T. Wylie, P. McInerney, K. G. Pillai, and R. A. Angryk, "On visualization techniques for solar data mining," *Astronomy and Computing*, vol. 10, pp. 32 – 42, 2015.

[36] D. Kempton, K. Pillai, and R. Angryk, "Iterative refinement of multiple targets tracking of solar events," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 36–44.

[37] J. M. Banda and R. A. Angryk. (2011) The solar dynamics observatory (sdo) content-based image-retrieval (cbir) system. [Online]. Available: http://cbsir.cs.montana.edu/sdocbir/php/index.php