Astronomy and Computing 13 (2015) 108–116

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Regional content-based image retrieval for solar images: Traditional versus modern methods

J.M. Banda *, R.A. Angryk

Georgia State University, 25 Park Place, Room 742, Atlanta, GA, USA

ARTICLE INFO

Article history:
Received 28 July 2015
Accepted 29 September 2015
Available online 23 October 2015

Keywords:
Image retrieval
Solar image analysis
Computer vision
Information retrieval
Content-based image retrieval
Large-scale retrieval
Big-data analysis

ABSTRACT

This work presents an extensive evaluation between conventional (distance-based) and modern (search-engine) information retrieval techniques in the context of finding similar Solar image regions within the Solar Dynamics Observatory (SDO) mission image repository. We compare pre-computed image descriptors (image features) extracted from the SDO mission images in two very different ways: (1) similarity retrieval using multiple distance-based metrics and (2) retrieval using Lucene, a general purpose scalable retrieval engine. By transforming image descriptors into histogram-like signatures and into Lucene-compatible text strings, we are able to effectively evaluate the retrieval capabilities of both methodologies. Using the image descriptors alongside a labeled image dataset, we present an extensive evaluation under the criteria of performance, scalability and retrieval precision of experimental retrieval systems in order to determine which implementation would be ideal for a production level system. In our analysis we performed key transformations to our sample datasets to properly evaluate rotation invariance and scalability. At the end of this work we conclude which technique is the most robust and would yield the best performing system after an extensive experimental evaluation, we also point out the strengths and weaknesses of each approach and theorize on potential improvements.

1. Introduction

With the launch of the Solar Dynamics Observatory (SDO) mission in 2012, solar physics entered the era of big data (Banda et al., 2014) in the first few months of the mission, and it is expected to produce more than twice the current amount of solar image data than all the previous solar image collecting missions have collected. Having collected over 50 million images, the SDO repository via a Content Based Image Retrieval (CBIR) system will allow researchers to query all indexed images with the purpose of finding other undiscovered images with similar characteristics. As part of the SDO Feature-Finding Team (SDO FFT) (Martens et al., 2012), the SDO CBIR system provides to the solar physics community the following advantages: (a) once new and unknown solar phenomena are discovered, our system allows users to look for similar images with occurrences of it through the SDO repository in a fast and efficient manner without the need to develop special feature-finding modules for this particular phenomena, (b) verification of existing feature-finding modules in order to determine their real accuracy when detecting solar phenomena in a strictly image similarity context, as we have shown in: Wan et al. (2014) (c) on a practical level, by currently training with known solar phenomena, our CBIR system allows researchers to find images that contain similar-looking phenomena to the ones present in the images they use to query our system not restricted only to the SDO mission (e.g. Transition Region and Coronal Explorer (TRACE) images).

The main focus of this work is to provide a comprehensive evaluation of the scalability of our existing SDO CBIR system (Website, 0000a; Banda et al., 2013a) from a larger-than-average system to a Big Data-capable tool that would allow researchers to query the entire SDO repository in an efficient and timely manner. The fundamentals of creating a fully scalable system capable of handling the growing SDO image repository have been introduced by Banda et al. (2014), but it was not until Banda and Angryk (2014b,a) that the first proof-of-concept system was presented. In this work we approach the issue of querying regions of interest found inside each image, a key component of state-of-the-art CBIR systems, and a feature, to our knowledge, not currently available in any solar CBIR systems. In order to address this key functionality, we propose pairing descriptor signatures, as introduced by Banda and Angryk (2014a), with the widely popular text-based search engine library Lucene (Banda and Angryk, 2014b). We provide a

* Corresponding author.
E-mail addresses: jbanda@gsu.edu (J.M. Banda), angryk@cs.gsu.edu (R.A. Angryk).

http://dx.doi.org/10.1016/j.ascom.2015.09.005
2213-1337/

comprehensive analysis of classical systems (traditional CBIR) and modern systems (Search-engine based) for solar images, under multiple test-scenarios allowing us to determine which type of system would be more suitable for public use in terms of satisfying researchers querying needs.

The overall organization of this paper is as follows: first, we provide the necessary background information in the following section, paying particular attention to the systems and datasets used. We then present our experimental analysis with added discussion. Lastly we provide our conclusions and conclude the paper with an outline future work.

2. Background information

In this section we prepare the reader with a brief overview of the SDO mission, the current state of the field of CBIR systems, as well as all the necessary information to make this paper self-sustainable in terms of the current SDO CBIR system, and datasets we experimented with. We also provide the necessary references for researchers that wish to acquire a deeper understanding into the design intricacies of the SDO CBIR system.

2.1. The solar dynamics observatory mission

As part of the Living With a Star program (Withbroe, 2000), the SDO spacecraft is found in geo-synchronous orbit continuously capturing full-disk images of the Sun for a period of 10 years. The purpose of the mission is to understand the mechanics of solar magnetic activity, from the generation of the solar magnetic field, to the release of magnetic energy in the solar wind, solar flares, etc. (Pesnell et al., 2012). SDO, to date, has already produced larger amounts of data than all previous solar data archives (Martens et al., 2012). The spacecraft carries three independent instruments: the Helioseismic and Magnetic Imager (HMI), which captures the motion of the sun's surface and measures the surface magnetic field (Scherrer et al., 2012) in full-disk images, the Extreme Ultraviolet Variability Experiment (EVE), which takes measurements (not images) of the spectral distribution of extreme ultraviolet radiation propagating from the sun (Woods et al., 2012), and the Atmospheric Imaging Assembly (AIA), which captures full-disk images of the sun in ten separate electromagnetic wavelength bands across the visual and ultra-violet spectrum (Lemen et al., 2012). The AIA images are the ones used in the SDO CBIR system. All SDO (and NASA) data has an open-access policy and is available to the general public through a variety of locations: http://sdo.gsfc.nasa.gov/data/ and http://www.helioviewer.org/.

2.2. Overview of content based image retrieval systems

Massive image repositories are becoming more readily available for science applications as new imaging technology and instruments evolve and produce more data. One of the main issues with such large image repositories is to be able to query them in an effective manner. Historically, the majority of widely-used image search engines have relied on comparing meta-data or textual tags associated with the images (Chakravarti and Meng, 2009), rather than the actual content included within the image file. Through the years, a small number of content-based image retrieval (CBIR) systems have facilitated general purpose querying tasks, such as Photobook (Pentland et al., 1996) from MIT which allows users to retrieve images by reducing said images to a smaller set of perceptually relevant coefficients and then computing basic similarities between them. Other systems are based on low-level image features, such as Candid (Ogle and Stonebraker, 1995), Chabot (Kelly et al., 1995) and QBIC (Flickner and Sawhney, 1995) from IBM. Note that the majority of the systems mentioned rely mostly on features such as shape, color, or texture for a complete-image similarity analysis, not for particular regions of interest within an image. Recent developments in the field have explored fuzzy approaches (Piedra-Fernandez et al., 2014), binning-strategies (Kamel et al., 2013), boosting methods (Kumar and Kumaraswamy, 2013), as well as provide more integrated tools (Hare et al., 2011) to aid the CBIR process.

Works describing region-based capabilities in CBIR systems started to appear with SIMPLIcity (James et al., 2001). Some of the earliest region-based capable systems were: Blobworld (Carson et al., 1999), Netra (Ma and Manjunath, 1999) and Walrus (Natsev et al., 1999), these systems besides offering full-image retrieval, had some type underlying region-of-interest based image representation allowing them to retrieve particular sections of the images. In Jing et al. (2003), the authors focused on adapting existing algorithms from CBIR into the region-based realm, but did not consider the scalability of their adaptations (Jing et al., 2004a,b; Li et al., 2000). This was mostly due to the fact that at the time, image repositories as large as Flickr, PhotoBucket, and SDO did not exist, or were not publicly available to the community. Currently, the issue of region-based image retrieval is still a complex problem since there are two main aspects to consider: (1) efficient image representation of newly created and more complex scientific image repositories is highly dependent of image domain, and (2) the scale in which the image data is being generated, ranging from several thousand images a day to terabytes of image data. The most relevant approaches and attempts of region-based image retrieval can be found in Huang et al. (2010) and Shrivastava and Tyagi (2014) showing great promise for color images. Unluckily, this does not generalize that well for SDO images as they are grayscale and the methods rely on color features to function properly. Other methods using clustering are very popular (Zakariya et al., 0000; Amory et al., 0000), but they have been shown to be optimal for datasets with very well defined regions of interest, such as a particular item or location in the image.

The majority of efficient and well-performing region-based retrieval representation approaches rely on arbitrarily constructed labels for sections of interest (Amores et al., 2007; Wei et al., 2012; Ri and Yao, 0000; Helala et al., 2012) but do not translate to images without clear objects to identify (such as solar image data). In the recent years, more robust and scalable systems like (Venkatraman and Kulkarni, 0000; Gu and Gao, 2012) are beginning to take advantage of Big Data technologies (Hadoop, MapReduce) for large-scale retrieval, but do not address the region-based querying issue. In terms of Lucene, LIRE (Lux and Chatzichristofis, 2008) proposes a Java library based on Lucene/Solr and a set of tools (Lux and Macstravic, 2014) for CBIR purposes. However, LIRE fails to deliver easy to adapt modules for other researchers to build their own custom image parameters into this system and be able to experiment without large amounts of code modifications to the original codebase. As a different approach to this problem, deep neural networks have been applied in the CBIR context in the past with very limited success (Wan et al., 2014). However, we expect that more researchers will turn to these methods in the future as they gain popularity within the image retrieval/classification communities.

2.3. Image parameters

Most of our earlier work focused on finding the ideal image parameters to extract from solar images of the SDO repository. In order to extract the image parameters, each 4096 × 4096 pixel image is segmented in 64-by-64 pixel cells, forming a 64 × 64 cell grid for each image. We then calculate ten numerical image