IMAGE RETRIEVAL ON COMPRESSED IMAGES: CAN WE TELL THE DIFFERENCE?

Juan M. Banda[1], Rafal A. Angryk[1], Michael A. Schuh[1], Petrus C. Martens[2 3]

[1]Dept. Computer Science, Montana State University, Bozeman, MT, USA
[2]Dept. Physics, Montana State University, Bozeman, MT, USA
[3]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA

ABSTRACT

In this work, we discuss the benefits of image compression on FITS image files to perform image retrieval tasks on the enormous NASA Solar Dynamics Observatory (SDO) image repository. With the objective of making solar image files more portable and easy to distribute and archive, we test several lossless compression algorithms as well as lossy compression algorithms in order to determine the rate we can compress standard FITS solar image files and still produce equal or comparable image processing and retrieval results. Our analysis comes from an image processing and retrieval viewpoints since we want to determine if the compression algorithms can reduce storage costs and analysis time. We believe that we might be able to hold huge repositories such as the SDO repository in a considerably smaller amount of disk space and still be able to perform the same image analysis experiments on this reduced and more portable repository.

Index Terms— Image compression, image processing, image retrieval, classification algorithms, data mining.

1. INTRODUCTION

Initially introduced in 1981 [1], the Flexible Image Transport System (FITS) became the de facto standard for image data in astronomy. Being designed for scientific data this format stores: photometric information, spatial calibration information, and plain-text lightweight ASCII header data among other science information. The format can also store other non-image data such as: photon lists, additional data cubes, and even multi-table databases. While this allows the format to be very flexible, it also makes it very bloated for fast image processing/retrieval operations that do not need all this additional science data. The format is also not ideal for practical storage purposes of massive repositories, when used for image retrieval tasks.

While developing a Content-Based Image-Retrieval (CBIR) system [2] for the latest NASA solar mission: Solar Dynamics Observatory (SDO) [3], we have encountered issues dealing the FITS image data files produced by the Atmospheric Imaging Assembly (AIA) instrument. AIA produces over 70,000 images a day for a total of 1.5 Terabytes of data using 10 different narrow wavelength pass bands. This data stream will continue over the next 10 years. One can easily see how this massive amount of data will quickly become prohibitive to have stored in one full central repository with all data available at all times. Sample images are shown in Fig. 1.

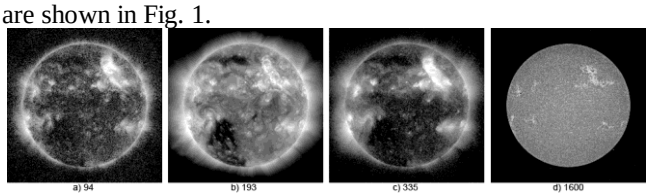


Figure 1. Sample SDO Images from the 94, 193, 335 and 1600 wavelengths for a similar time range

We are fully aware that some Solar scientists have a use for the extra data in the FITS image data files like the headers and the photon counts, however, for effective image retrieval we need smaller image files that can be processed with ease and speed. Considering that original uncompressed FITS files for the SDO mission are each 64 MB in size (4K x 4K resolution), one can immediately see the advantages of working with a considerably smaller size in terms of processing time and portability of the dataset. We are sure that the image processing community would be thrilled to analyze a massive image repository as the SDO dataset, but its current size and lack of portability make it nearly impossible for researchers to have a complete copy of it. There are significant efforts in the scientific community to offer reduced version of this dataset, such as the Helioviewer portal [4] that allows the download of single images in JP2 format, but to our knowledge there are no other smaller versions of this dataset or even labeled subsets available from it until [5]. In our work we present the argument that for feature extraction and image retrieval, we do not need to have the original FITS image data files present to be able to perform similar tasks on compressed image format. The amount of compression achieved and the potential trade-off is also presented to the readers to let them make a decision of what better suits their quality thresholds.

The rest of the paper is organized in the following way: Sec. 2 we introduce background information, such as the datasets used, the compression algorithms we utilized, how we created our feature vectors and the classification algorithms uses for our analysis. Sec. 3 includes a description and commentary of our similarity evaluation between feature vectors. In Sec. 4 we present our classification evaluation

performed on the image data. Finally our conclusions are outlined in Sec. 5.

2. BACKGROUND INFORMATION

Since this work is a product of our main goal of building a large-scale CBIR system that allows users to find similar images of solar events (flares, sigmoids, etc), we will list the individual references of other publications that address in detail the reason behind the choices we have made for image parameters, our feature vector generation, and classification algorithm selection, since all of these discussions are beyond the scope of this short paper.

2.1. Datasets

In order to provide a comprehensive comparison of the compression algorithms, we selected two different solar datasets which have FITS files as their original source. To gain the interest of computer scientists not working with solar data, we have also included two medical image datasets that have been used in our previous work [6] to show their similarities to solar image in terms of image retrieval performance based on our image parameter selection for the solar datasets.

2.1.1 SDO Dataset

Taken from January 1st, 2013 to January 31st, 2013, we selected images that had solar event labels provided by other modules part of the Feature Finding Team [7] that is in charge of automatically identifying solar phenomena events on SDO data. All these event labels can be in the Heliophysics Events Knowledge (HEK) Base found at [8].

This dataset is comprised of 2,512 FITS image data files with a resolution of 4,096x4,096 pixels and taken from 3 different wavelengths; note that each image was normalized following the standards set by Helioviewer for image display. The dataset features 2,766 labeled events each from five different classes of solar events: Active Regions, Sigmoids, Coronal Holes, Flares and Quiet Sun. It can be found for download in pre-processed form (image parameters only, for now) at [9]. Sample images can be seen in Fig.1.

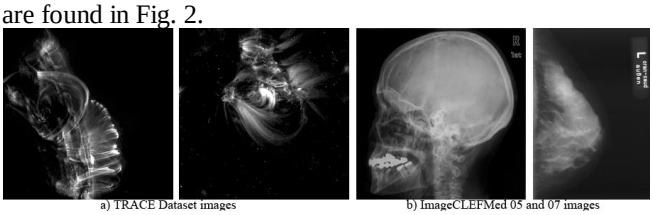
2.1.2 TRACE Dataset

This dataset is comprised of 1,600 FITS image data files with a resolution of 1,024x1,024 pixels and was first introduced in [10]. The dataset is divided in 8 balanced classes (200 images per class) representing the following solar events: Active Region, Coronal Jet, Emerging Flux, Filament, Filament Activation, Filament Eruption, Flare and Oscillation. The benchmark dataset both in its original and pre-processed format is freely available to the public via [11]. These images were taken from the TRACE mission, which was launched by NASA several years before the SDO mission. Sample images are featured in Fig. 2.

2.1.3 ImageCLEFMed05 and 07 – Medical Datasets

These datasets are both comprised of 1,600 PNG image files with a varying resolution, but all including one side of at least 512 pixels in size. The images in this dataset feature radiographs of different ages, genders, view positions and pathologies. We randomly selected 1,600 images from the 2005 and 2007 datasets balancing eight classes with 200 images each.

ImageCLEF is the cross-language image retrieval track which is run as part of the Cross Language Evaluation Forum (CLEF) [12]. In their medical retrieval task, we find several datasets available from 2005 to 2012. Sample images are found in Fig. 2.



a) TRACE Dataset images b) ImageCLEFMed 05 and 07 images

Figure 2. Sample images from the TRACE and ImageCLEFMED datasets

2.2. Compression Algorithms

To provide a comprehensive analysis of the most popular compression algorithms, we used 6 different lossless algorithms and four different compression settings on the lossy JPEG compression algorithm [13]. Table 1 presents the algorithms used and their settings.

Compression Algorithm	Label	Notes
Flexible Image Transport System [1]	FITS	Lossless - Default settings
Tagged Image File Format [14]	TIF	Lossless - Default settings
Graphics Interchange Format [15]	GIF	Lossless - Default settings
Portable Network Graphics [16]	PNG	Lossless - Default settings
Joint Photographic Experts Group (JPEG) [13]	JPG	Lossless specification
JPEG 10%	JPG10	Lossy - Quality (Q) = 10%
JPEG 25%	JPG25	Lossy - Quality (Q) = 25%
JPEG 50%	JPG50	Lossy - Quality (Q) = 50%
JPEG 75%	JPG75	Lossy - Quality (Q) = 75%
JPEG 2000	JP2	Lossless specification

Table 1. Compression Algorithms used, their labels (for experiments), and their settings.

NOTE: We based our selection of algorithms due to their popularity in literature [17, 18, 19, 20] and their availability in image processing packages, in particular ImageMagick [21] and OpenCV [22] that have been used for this analysis.

2.3. Feature Vector Generation

In order to represent our images in numeric vectors for our image retrieval tasks, we first segment our solar images in 128x128 pixel cells (i.e. 64x64 grid for the SDO dataset and 8x8 grid for the TRACE dataset), since the image segmentation technique has shown good results in previous work [10, 23]. We then calculate ten image features for each individual cell, thus being left with a highly dimensional feature vector. The ten image parameters [24] that we have concluded to be the most useful for our task are: the Mean

intensity, the Standard Deviation of the intensity, the Third Moment and Fourth Moment, Uniformity, Entropy, and Relative Smoothness, Fractal Dimension [25], and two Tamura texture attributes: Contrast and Directionality [26]. This process is shown in Fig. 3.

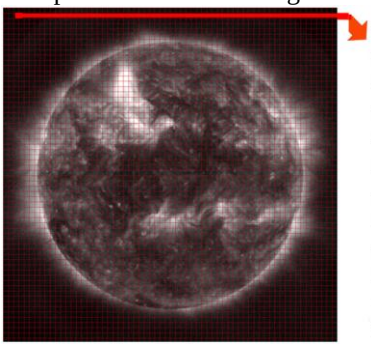


Image parameter	Value
Entropy	0.1521
Mean	0.2695
Standard Deviation	0.5751
3rd Moment (skewness)	0.3652
4th Moment (kurtosis)	0.1112
Uniformity	0.9984
Relative Smoothness (RS)	0.2154
Fractal Dimension	0.0014
Tamura Directionality	0.1512
Tamura Contrast	0.1412

Figure 3. Example of feature extraction on SDO image cells

For the ImageCLEFMed datasets we applied the 8x8 grid with each cell being proportional to each image size.

2.4. Classification Algorithms

As in our previous work [6, 24, 27], we have opted to present our classification evaluation using four different classifiers: Naïve Bayes (NB) [28] and Support Vector Machines (SVM) [29] using a linear kernel function. Since both of the previously mentioned classifiers are linear, we opted to use two more decision tree classifiers: C4.5 [30] and Random Forest (RF) [31]. The justification for our classifier selection is out of the scope of this work and has been greatly explained in [6, 32].

3. FEATURE VECTOR SIMILARITY EVALUATION

For this part of our analysis, we compare the feature vector extracted from the original image formats (FITS for SDO and TRACE datasets and PNG for the ImageCLEFMed datasets) with the feature vectors extracted from the compressed image files of the same dataset. This comparison is performed as in (1), allowing us to determine the magnitude of the difference between vectors.

$$Diff = \left(1 - \frac{\sum_{r=1}^{N}\sum_{c=1}^{N}\sum_{p=1}^{10}\left|O_{(r,c),p}\right|}{\sum_{r=1}^{N}\sum_{c=1}^{N}\sum_{p=1}^{10}\left|O_{(r,c),p} - C_{(r,c),p}\right|}\right) * 100\% \qquad (1)$$

where p indicates the number of parameters and r and c indicate the corresponding cell row and column number. N denotes the total number of rows and columns in our grid since it varies between the SDO images (64x64) to the rest of the images (8x8), due to the images resolutions.

	FITS	TIF	GIF	PNG	JPG	JPG10	JPG25	JPG50	JPG75	JP2
SDO	0%	0.2%	0.3%	0.3%	0.1%	2.8%	1.5%	1.5%	0.5%	1.1%
TRACE	0%	0%	0%	0%	0.2%	6.2%	4.1%	3.8%	2.3%	0.2%
Med05	NA	3.1%	0%	0%	0%	2.8%	2.8%	2.8%	2.6%	2.8%
Med07	NA	0.5%	0.4%	0%	0.4%	0.2%	0.6%	0.6%	0.6%	0%

Table 2. Difference between original and compressed feature vectors

As we can see in Tab. 2, for the SDO dataset, the difference between the feature vectors of the FITS image parameters to the compressed images parameters is very minimal, with most of the lossless methods coming to be less than 0.4% and JP2 coming in as the 'worst' with 1.1%. Even the lossy methods are still within reasonable differences with 2.8% for the highly compressed JPG10. Very similar trends follow for the rest of the datasets, with TRACE and JPG10 finding the biggest difference between vectors with a very acceptable 6.2%. In Fig. 4 we will show a plot of the differences between the image parameter vectors for the SDO dataset, this plot will more clearly illustrate the differences.

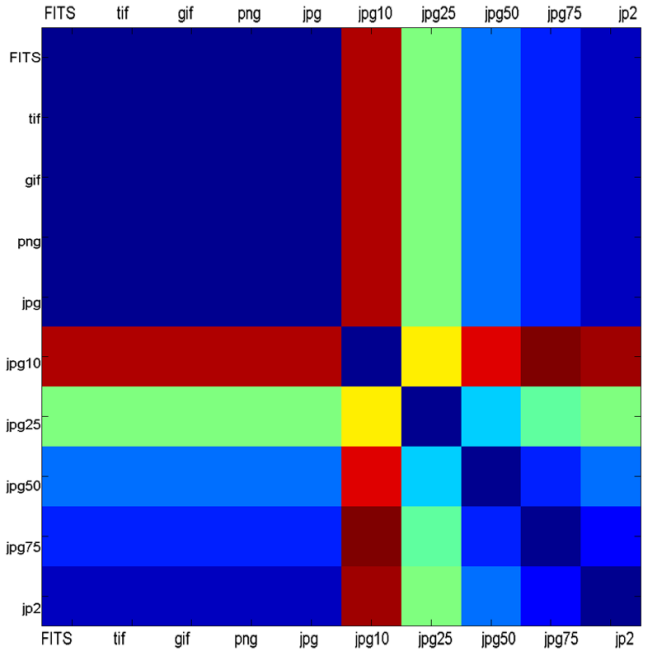


Figure 4. Plot of the difference between feature vectors for the SDO dataset. Note: Values closer to RED indicate bigger difference between vectors. The darker the BLUE color is, indicates high similarity

As we can clearly see in the figure, all the lossless algorithms produce nearly identical feature vectors with a very minimal exception for the JP2 algorithm as seen on Tab.2 . This is indicated within the darker cluster for FITS, TIF, GIF, PNG. The vectors differ the most for the JPEG10, which is the most compressed format we have, this figure also indicates how their quality deteriorates consistently as we compare the compressed vectors between themselves.

We present in Fig 5. plots that show how the feature vectors differ for the TRACE, and CLEFMed datasets. These plots are less interesting due to the fact that the images have a lower resolution than SDO ones, allowing for bigger changes in the image parameters data values thanks to more evident compression artifacts in the lossy methods for most of the datasets. For the lossless algorithms we have very similar behavior as for the SDO dataset. This is clearly shown with the two different colored clusters found.

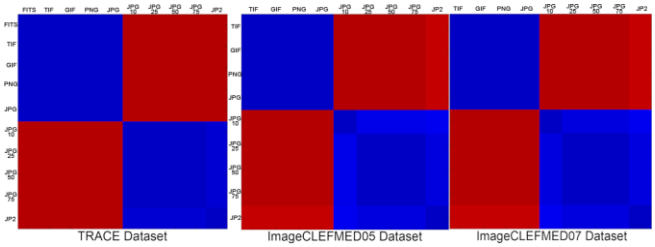


Figure 5. Plot of the difference between feature vectors for the TRACE and Medical datasets

These results are indeed very surprising since they lead us to believe that even when working when compressed images, the selected image parameters will perform very similarly and we will try to show this on the next section.

4. CLASSIFIER-BASED EVALUATION

During our classification evaluation, we ran our four classification algorithms using 10-fold cross-validation. We used the WEKA [33] implementations of the four algorithms. As a side note, we are not trying to find the best results here, we are just showing how the in theory degraded parameter values of compressed images would influence distinguishing between image labels. Fig. 6 presents the classification accuracy results of the labeled images of the SDO Dataset.

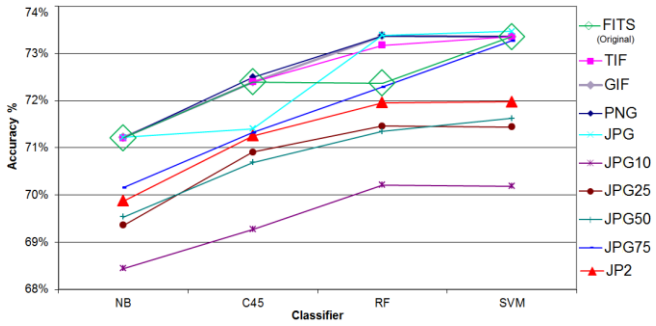


Figure 6. Classification accuracy results for the SDO Dataset

Similarly to the results from feature vectors differences, the classification accuracy results show the same algorithms performing equally good in the classification tasks. The highly compressed images (JPEG75 to JPEG10) do show degraded performance, but they are still within 3% accuracy from the original data. This results show that in our particular approach to finding/identifying solar phenomena, we can use compressed images without many issues. The level of compression that one might select is closely tied to how much accuracy one is willing to sacrifice.

On Fig. 7, we show the classification results for all the datasets we used. TRACE shows the same tendencies as SDO. And the other datasets that have lower resolution and PNG original image files have the interesting pattern of not diverging much from the original classification values, allowing us to believe that image compression can also be performed on these types of datasets without much loss in accuracy.

The results found in the classification evaluation seem to validate our vector difference results providing very similar finding and making us believe we can generalize them in an image retrieval context fairly easily.

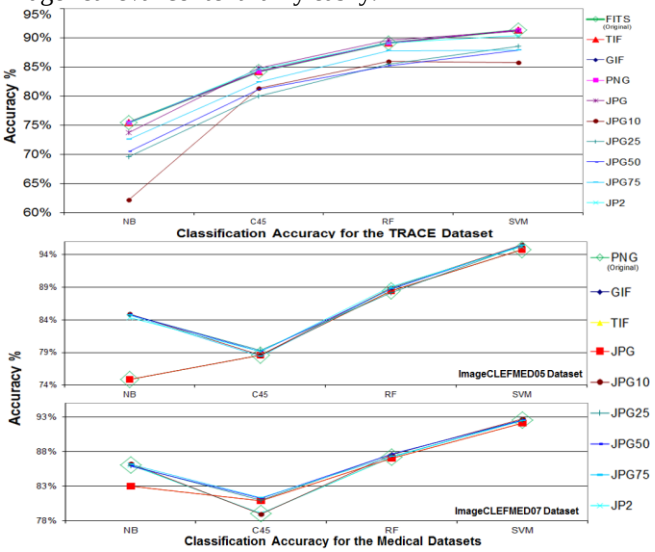


Figure 7. Classification accuracy for the TRACE and Medical datasets

5. CONCLUSIONS

As we shown in the previous analyses, we can safely say that we can compress our image data and still maintain a similar feature vector structure that will allow us to perform classification tasks properly and saves us a considerable amount of storage resources. In order to really see how much can we save in storage costs we present table 3 that indicates space savings and compression ratios for the SDO dataset

Format	Space Savings*	Compression Ratio**	Accuracy loss ***
FITS	0.0%	100/100	0.1%
TIF	50.0%	50/100	0.3%
GIF	96.9%	3/100	0.4%
PNG	75.2%	25/100	0.4%
JPG	74.8%	25/100	0.2%
JPG10	99.2%	1/100	2.5%
JPG25	98.4%	2/100	1.2%
JPG50	97.2%	3/100	1.3%
JPG75	95.1%	5/100	0.7%
JP2	79.7%	20/100	1.3%

Table 3. Space savings and compression ratio for SDO images.
* Calculated as: (1 – (uncompressed size / compressed size))*100%
** Calculated as: compressed size / uncompressed size
*** Calculated as: original accuracy – compressed accuracy

As one can clearly see, the tradeoff for over 95% compression is very little in the tests performed on our image parameters for the SDO dataset. This leads us to believe that we can highly compress our original FITS images and still achieve decent classification results. While we do not want to go as far as using JPG10, we can reduce our storage by 75% using lossless PNG and JPG algorithms and lose a very minimal amount of performance.

While we have shown very good results on our image parameters on the compressed images, we theorize that these results are closely tied with the particular types of image

parameters we are using (textural) and the peculiar nature of the images we have (greyscale, fuzzy objects – in the solar images). These results will probably not generalize properly for color images that are more prone to compression artifacts than our images. We also show that we can compress the already small medical images and lose very little performance, leading us to believe that we might be able to get similar performance while also reducing the pixel size of our Solar images, an analysis that will be left for future work. All data and WEKA files to reproduce our results are available at [9].

6. ACKNOWLEDGEMENTS

This work was supported in part by two NASA Grant Awards:
1) No. NNX09AB03G, and 2) No. NNX11AM13A. Any opinions expressed herein are those of the authors and do not necessarily represent the views of the National Aeronautics and Space Administration (NASA).

7. REFERENCES

[1] D.C. Wells, E.W. Greisen and R.H. Harten, "FITS - a Flexible Image Transport System", Astronomy and Astrophysics Supplement, Vol. 44, P. 363, 1981.

[2] J. M. Banda, R. Angryk, P. Martens. "On dimensionality reduction for indexing and retrieval of large-scale solar image data". Solar Physics: Image processing in the petabyte era. Springer 2012. DOI:10.1007/s11207-012-0027-4.

[3] Solar Dynamics Observatory [Online], Available: http://sdo.gsfc.nasa.gov/. [Accessed: January 22nd, 2013]

[4] Helioviewer [Online], Available: http://www.helioviewer.org/. [Accessed: January 22nd, 2013]

[5] M.A. Schuh, R.A. Angryk, K.G. Pillai, J.M. Banda, P. C. Martens. "A Large-Scale Solar Image Dataset With Labeled Event Regions" . In Proc. of the International Conference on Image Processing (ICIP) (2013), pp. 4349–4353.

[6] J.M. Banda, R. Angryk, P. Martens. "On the surprisingly accurate transfer of image parameters between medical and solar images", ICIP-IEEE '11, Brussels, Belgium, September 2011, pp. 3730-3733.

[7] P. C. H. Martens, G. D. R. Attrill, A. R. Davey, *et al*., "Computer Vision for the Solar Dynamics Observatory (SDO)". Solar Physics: 275:79-113, 2012.

[8] Heliophysics Event Registry [Online] Available: http://www.lmsal.com/~cheung/hpkb/index.html [Accessed: January 22nd, 2013]

[9] IPTA 2014 supplemental website [Online]. http://www.jmbanda.com/IPTA2014/. [Accessed: January 22nd, 2013]

[10] J.M Banda and R. Angryk "An Experimental Evaluation of Popular Image Parameters for Monochromatic Solar Image categorization" Proceedings of the 23rd international Florida Artificial Intelligence Research Society conference (FLAIRS-23), Daytona Beach, Florida, USA. pp. 380-385. 2010.

[11] SDO Dataset (MSU) [Online], Available: http://www.cs.montana.edu/angryk/SDO/data/ [Accessed: January 22nd, 2013]

[12] Cross Language Evaluation Forum [Online], Available: http://www.clef-campaign.org/ [Accessed: January 22nd, 2013]

[13] W. B. Pennebaker and J. L. Mitchell. *JPEG still image data compression standard (3rd ed.)*. Springer. 1993.

[14] Tagged Image File Format RFC3949 (latest specification). [Online], Available: http://tools.ietf.org/html/rfc3949/ [Accessed: January 22nd, 2013]

[15] Graphics Interchange Format, Version 87a. [Online], Available: http://www.w3.org/Graphics/GIF/spec-gif87.txt [Accessed: January 22nd, 2013]

[16] ISO/IEC 15948:2004 - Portable Network Graphics (PNG): Functional specification. [Online], Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=29581 [Accessed: January 22nd, 2013]

[17] G. Schaefer. "Content-based retrieval of compressed images". In International Workshop on DAtabases, TExts, Specifications and Objects. pp. 175-185. 2010.

[18] R.F Chang, J.W. Kuo and H.C. Tsai. "Image retrieval on uncompressed and compressed domains". In Image Processing, Vol. 2, pp. 546-549. 2000.

[19] D. Cerra and M. Datcu, "Image retrieval using compression-based techniques," Source and Channel Coding (SCC), 2010 International ITG Conference pp. 18-21.2010

[20] G. Schaefer. "Does compression affect image retrieval performance?." International Journal of Imaging Systems and Technology. Vol. 18.2-3 pp. 101-112. 2008.

[21] ImageMagick Software Suite. [Online], Available: http://www.imagemagick.org/. [Accessed: January 22nd, 2013]

[22] Open Source Computer Vision Library. [Online], Available: http://opencv.org/. [Accessed: January 22nd, 2013]

[23] J.M Banda and R. Angryk "On the effectiveness of fuzzy clustering as a data discretization technique for large-scale classification of solar images" Proceedings of the 18th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '09), Jeju Island, Korea, pp. 2019-2024 August 2009

[24] J.M Banda and R. Angryk "Selection of Image Parameters as the First Step Towards Creating a CBIR System for the Solar Dynamics Observatory". International Conference on Digital Image Computing: Techniques and Applications (DICTA). Sydney, Australia, 2010.

[25] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. New York: W. H. Freeman, pp. 41-45, 1991.

[26] H. Tamura, S. Mori, T. Yamawaki. "Textural Features Corresponding to Visual Perception". IEEE Transaction on Systems, Man, and Cybernetics 8(6): pp. 460–472. 1978.

[27] J.M. Banda, R. Angryk, P. Martens, "Quantitative Comparison of Linear and Non-linear Dimensionality Reduction Techniques for Solar Image Archives", Proceedings of the 25th International FLAIRS Conference (FLAIRS-25), Marco Island, Florida, pp. 376-381, 2012.

[28] M. Minsky. "Steps toward Artificial Intelligence." Proceedings of the IRE 49(1):8-30, 1961.

[29] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag. 1995.

[30] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann. 1993.

[31] L. Breiman. "Random Forests". Machine Learning 45 pp. 5–32. 2001.

[32] J.M. Banda, R. A. Angryk, P. C. H. Martens. "imageFARMER – Introducing a framework for the creation of large-scale content-based image retrieval systems". International Journal of Computer Applications 79(13):8-13, October 2013. Published by Foundation of Computer Science, New York, USA

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. "The WEKA Data Mining Software: An Update" SIGKDD Explorations, Volume 11, Issue 1, 2009