Steps Toward Large-scale Solar Image Data Analysis to Differentiate Solar Phenomena

J.M. Banda[1] · R. A. Angryk[1] · P.C.H. Martens[2,3]

Abstract We detail the investigation of the first application of several dissimilarity measures for large-scale solar image data analysis. Using a solar-domain-specific benchmark dataset that contains multiple types of phenomena, we analyzed combinations of image parameters with different dissimilarity measures in order to determine which combinations will allow us to differentiate among the multiple solar phenomena from both intra-class and inter-class perspectives, where by class we refer to same types of solar phenomena. We also investigate the issue of reducing data dimensionality by applying multidimensional scaling to the dissimilarity matrices we produced using the previously mentioned combinations. As an early investigation into dimensionality reduction, by applying multidimensional scaling (MDS) we will investigate how many MDS components are needed in order to maintain a good representation of our data (in a new artificial data space) and how many can be discarded in order to enhance our querying performance. Finally, we present a comparative analysis among several classifiers in order to determine the quality of the dimensionality reduction achieved with the aforementioned combination of image parameters, similarity measures, and multidimensional scaling (MDS).

Keywords: Solar image data analysis, content-based image retrieval (CBIR), dissimilarity measures

1. Introduction

In this article we present some of our preliminary steps toward the ambitious goal of building a Content Based Image Retrieval (CBIR) system for the *Solar Dynamics Observatory* mission (SDO: sdo.gsfc.nasa.gov/). Our work is motivated by the recognition that the massive amount of data that the SDO mission is transmitting presents a heretofore seldom-addressed problem in terms of image analysis for scientific purposes. With SDO's *Atmospheric Imaging Assembly* (AIA) alone, the mission will generate eight 4096 pixels × 4096 pixels images every ten seconds, leading to a data transmission rate of approximately 700 gigabytes per day (the entire mission is expected to send about 1.5 terabytes of data per day, for a minimum of five years). Hand labeling of these images will be impossible and backtracking to find similar images with new and currently undiscovered phenomena will be a nearly impossible task without the aid of CBIR technology. Such a CBIR system will allow researchers to query the indexed images of the SDO repository for images similar to the ones they are currently analysis. By using different similarity measures, researchers will be presented with different sets of results that might satisfy their queries better than using the standard Euclidean distance. With this CBIR system we will be able to provide the following advantages to the solar physics community: i) As mentioned earlier, once a new and unknown phenomenon is discovered, our system will allow users to look for similar occurrences of it in the SDO repository in a very fast and efficient manner without the need to develop a new feature-finding module for this particular phenomena, ii) verification of existing feature-finding modules in order to determine their real accuracy when detecting solar phenomena in a strictly image similarity context, iii) on a practical level, based on the training with current known solar phenomena, our CBIR system will allow researchers to find images that contain similar-looking phenomena to the ones present in the

[1]Montana State University, Department of Computer Science, Bozeman, MT USA {juan.banda,angryk}@cs.montana.,edu
[2] Department of Physics, Montana State University, 247 EPS, Bozeman, MT 59717-3889, USA martens@physics.montana.edu
[3] Harvard Smithsonian Center for Astrophysics. 60 Garden Street, Cambridge, MA. USA

images they use to query our system that are not restricted only to the SDO mission (e.g. *Transition Region and Coronal Explorer* (TRACE) images).
One of the main contributions of our work is to determine similarity (or dissimilarity), an issue that in the literature has always been presented as a very domain-specific problem that needs to be addressed in the context of solar data before we can continue to build our CBIR system. One can find preferences for different dissimilarity measures used in different information-retrieval (IR) systems, and we tried to cover most of them in this work. For instance, the majority of text-based IR systems use the cosine dissimilarity measure (Tan, Steinbach, and Kumar, 2005), while CBIR systems for color images often use the Kullback–Leibler Divergence (KLD) measure (Deselaers, Keysers, and Ney, 2008). CBIR community available tools like Lire (Lux and Savvas, 2008), do not allow this testing flexibility and are particularly designed for natural scene images, the biggest focus of the CBIR community, and modifying a tool like this might turn out to be a bigger task than creating a custom fitted one for solar data. Of all the successful non-proprietary CBIR systems that we investigated in medical (Deselaers, Keysers, and Ney, 2008) and other domains (Datta, Li, and Wang, 2005); none of them have dealt with solar data or the volume of data that the SDO mission generates. The only two systems to our knowledge capable of dealing with several millions of images are Google's "Similar Images" and TinEye. "Similar Images" was an experimental system that in 2009 was incorporated into the regular Google Image search becoming proprietary-technology. The same applies to TinEye, making both systems (and details) unavailable for research purposes. While both are freely available for use, we doubt that NASA or any solar physicist will be willing to upload 1.5 Terabytes of data a day to their servers in order to try to make use of them, we also doubt that these companies will be willing to comply with such a daunting task for a niche application (solar image analysis *vs*. general image analysis).

1.1 This Work in Context of the Future SDO CBIR System

In Figure 1 we present how a traditional CBIR system is designed and queried by a user. In this figure we also highlight in red the sections of the system and query mechanism that this work covers. While the complete system is not fully available and designed as of now, each of these components needs to be properly analyzed before making final implementation decisions. Using the image dataset (Image Collection on Figure 1) that we have introduced in Banda and Angryk (2010 a), we proceed to segment our images (Image Segmentation on Figure 1) and extract our image parameters (Image Parameters Extraction on Figure 1). More details of the parameters selected and the segmentation method are presented in Banda and Angryk (2010 a) however, on section 2.2.1 we show how these statistical image parameters are calculated. Said parameters are then transformed into feature vectors (Feature Vectors Creation in Figure 1) for a more compact and reduced representation, we outline this process on section 2.2.2. At this stage, we are now confronted with the problem of determining the most informative dissimilarity measures for our solar images strictly based on the image parameters that we have extracted. This work presents our experimentation toward achieving sections Feature Vector Indexing and Similarity Comparison and Retrieval in Figure 1. To achieve this, we experimented with 18 similarity measures that are widely used for clustering, classification, and retrieval of images (Ojala, Pietikainen, and Harwood, 1996; Lam *et al.*, 2000; Aggarwal,

Hinneburg, and Keim, 2001; Guo *et al.*, 2002; Francois, Wertz, and Verleysen, 2005) in order to determine which ones would provide a better differentiation of the solar phenomena presented on our image dataset. In order to determine which combination of image parameters and similarity measures work best, we investigated over 180 combinations. These experiments were performed to help identify the most (and least) informative and useful combinations for our system similarity comparison, retrieval, and indexing needs.

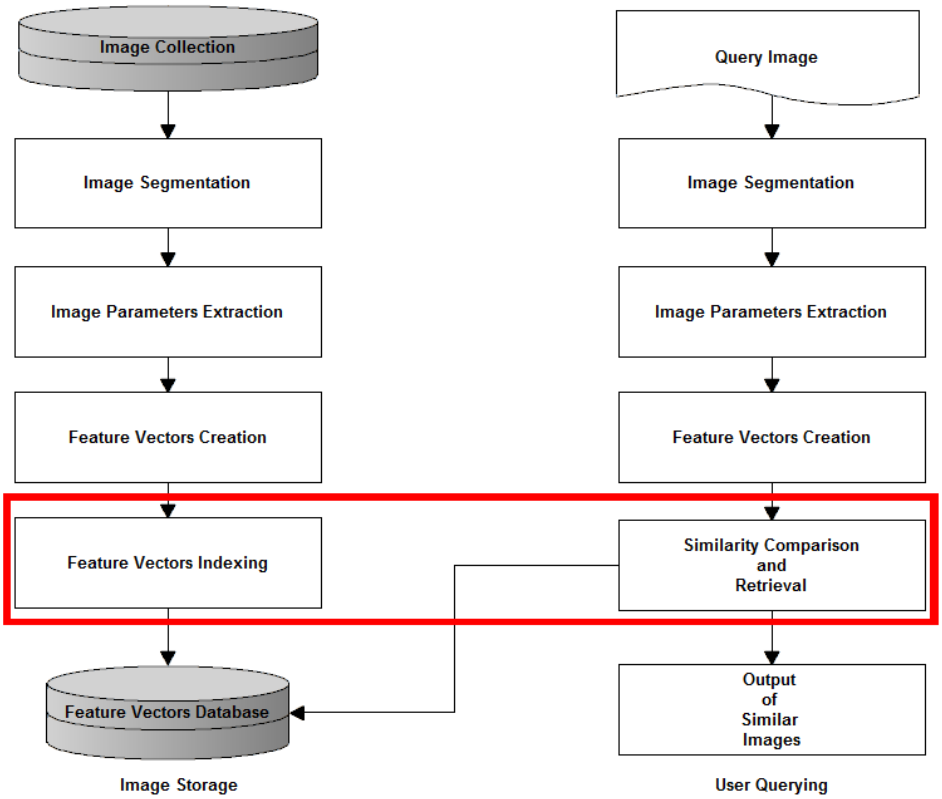


Figure 1. Traditional elements of a CBIR system. Highlighted components are being addressed in this work.

Besides qualitatively determining which combinations of dissimilarity measures and image parameters work best via the visual analysis of dissimilarity matrices (using scaled image plots), we also performed extensive quantitative analyses by applying multidimensional scaling (MDS) to our resulting dissimilarity matrices (taking advantage of the different characteristics highlighted by each individual measure) and then compared the resulting reduced dimensional space representations of our feature vectors with three different classification algorithms in order to mimic our future retrieval component to be used on our CBIR system. This MDS method has widely been used for visualization and dimensionality reduction by researchers in different areas for image processing and retrieval (Beatty and Manjunath, 1997; Rubner, Guibas, and Tomasi, 1997; Borg and Groenen, 2005; Datta, Li, and Wang, 2005). By applying MDS to our dissimilarity matrices, we: i) provide a mechanism for the construction of a 2D or 3D visualization of our dataset's dissimilarities that depicts class (different types of solar phenomena) separation in a convenient way, ii) estimate the amount of dimensionality reduction that we can achieve with our data points mapped into a new artificial dimensional space generated by MDS, and evaluate any performance costs by presenting a comparative evaluation using several classification algorithms.

In order to measure the quality of our combinations of dissimilarity-measures and image-parameters, and our dimensionality-reduction estimation methodology we set up two different ways of limiting the number of MDS components. We quantitatively evaluate our work using comparative analysis, where we compare the two different component-selection methods by presenting classification results for three different classifiers. This allowed us to determine how to select our MDS components in order to achieve similar (or even better) classification results than with our original data. In our particular CBIR system, with the expected growth of our repository, the applicability of dimensionality reduction is very important in terms of allowing us to reduce our query performance. Based on the proposed grid-based representation — where we extract image parameters from 128 × 128 pixel cells from the solar images — we would have to store one 10240 dimensional vector per image, resulting in a total of 5.27 Gigabytes of text data per day, which will be added to our CBIR system's database. Indexing and querying a database of this size and growth is a challenge on its own, and dimensionality reduction will greatly help us in providing an efficient and responsive CBIR system for the community to use.

We present this work to the broader community of solar physicists and computer scientists not only to contribute to the existing knowledge on solar data analysis (Banda and Angryk, 2009; 2010 a, b, c), but also to obtain valuable feedback from the community. The potential feedback from solar physicists using image parameters different than those presented in this work will be especially valuable. We look forward to building new collaborations with domain experts who are working on identifying individual solar phenomenon, as we intend to proceed with additions to our previously published benchmark dataset (www.cs.montana.edu/angryk/SDO/data/). Since the SDO mission has been launched, the need to accurately detect and classify different types of solar phenomena in an automated way is of vital importance. We are open to discussion and would greatly appreciate any feedback offered.

With the outline workflow presented here, other solar physicists can greatly benefit from knowing which dissimilarity-measure – image-parameter combinations work well, and therefore can improve on their own work in the domain of classification of specific solar phenomenon. As we mentioned in Banda and Angryk (2010 a), the results are very specific to the domain of individual solar phenomenon. They allow researchers who are working on a particular type of solar event (e.g. flares) to use a combination of dissimilarity-measure – image-parameter measures that better serve their classification purposes.

This article is organized in the following way: We present the necessary methodology in Section 2. In Section 3 we present our experiments and results. Section 4 contains our overall conclusions based on these experimental results, and Section 5 describes future work.

2. Methodology

In this section we identify all of the components that are needed for our experimental evaluation. We first characterize the benchmark dataset of images that we will be using for our experiments. After the dataset has been introduced, we will detail how we extract numerical image parameters from the dataset images and transform them into feature vectors for similarity analysis by explaining our normalization and data preprocessing. We then explain the dissimilarity measures that we will be using in our work as well as the principles

of Multi-Dimensional Scaling (MDS) and how we are using this method to visualize our data relationships and make a preliminary estimate of dimensionality reduction in a very introductory manner and it is not part of the actual system. Lastly, we explain the classification algorithms we utilized to provide a quantitative comparison of the intricacies of our proposed experimental evaluation.

2.1 Benchmark Dataset Used

Our dataset, first introduced in Banda and Angryk (2010 a), consists of 1600 images obtained in January of 2008 from the NASA TRACE (Handy *et al.*, 1999) mission via the *Heliophysics Event Knowledge Base* (HEK) (www.lmsal.com/~cheung/hpkb/index.html). Our dataset is divided in eight equally balanced classes representing eight types of solar phenomena. Table 1 enumerates these eight types of solar phenomena.

Table 1. Characteristics of Our Benchmark Dataset

Phenomenon/Event Name	# of images	Wavelength [Ångström]
Active Region	200	1600
Coronal Jet	200	171
Emerging Flux	200	1600
Filament	200	171
Filament Activation	200	171
Filament Eruption	200	171
Flare	200	171
Oscillation	200	171

The benchmark dataset, both in its original and pre-processed format, is freely available to the public via Montana State University's server (www.cs.montana.edu/angryk/SDO/data/). All of our TRACE images in the dataset have been annotated for events and phenomenon by solar physicists who each have detailed knowledge of both the TRACE observatory and the type of solar imagery involved, and reported their findings in the HEK (www.lmsal.com/~cheung/hpkb/index.html). In this work we assign one phenomenon label per image regardless of the event location in the image.

2.2. Feature Vectors Generation

In order to build the feature vector representation of our images, first we need to extract information about their content. For this task we use image parameters that are designed to calculate particular signatures based on the image (or region) texture, gray-levels distribution, and produce different numerical values to populate our feature vector representation.

2.2.1 Image Parameter Extraction

Based on our literature review, we decided to use some of the most popular image parameters used in other domains such as: medical image analysis, text recognition, natural scene image analysis, and traffic image analysis (Pentland, 1984; Chaudhuri and Nirupam, 1995; Cernadas *et al.*, 2005; Wen-lun, Zhong-ke, and Jian, 2005; Holalu and Arumugam, 2006; Deselaers, Keysers, and Ney, 2008; Devendran, Hemalatha, and Amitabh, 2009). Since the aforementioned respective image parameterizations have shown to be very domain-specific, we performed

our own investigation on the evaluation of these image parameters. All parameters where obtained from Gonzalez and Woods (2006), with the exception of the fractal dimension (Schroeder, 1991) and the Tamura textural parameters (Tamura, Mori, and Yamawaki, 1977).

The ten image parameters that we used for this work are presented in Table 2. Note that these image parameters are not exhaustive, and there are many other parameters that we could have tested. In our previous works, we started with a larger list of parameters, but we have since discarded some based computational expense, performance, and relevance (Banda and Angryk, 2009; and 2010 a).

Table 2. Extracted image parameters. For the grey-scale image (cell) z, $p(z_i)$ denotes the histogram count value for the i^{th} gray level in our image (cell) as described by Gonzalez and Woods (2006). In our case we have 255 different grey levels. Correspondingly, z_i will be the i^{th} grey level for P1, P4, P5, and P10. For P3, P6, and P7, z_j indicates the value of each pixel from the image (cell), said j will cover all values in said image (segment) up until K, where K is 128 × 128 = 16384 number of pixels. The P2 is calculated based on the box-counting method where N(ε) is the number of boxes of side length ε required to cover the image cell. In P6 we used σ^2, which indicates the variance, as defined for P7.

Label	Image parameter	Formula	
P1	Entropy	$E = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$	(1)
P2	Fractal Dimension	$D_0 = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \frac{1}{\epsilon}}$	(2)
P3	Mean	$m = \frac{1}{K} \sum_{j=1}^{K} z_j$	(3)
P4	3rd Moment (Skewness)	$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$	(4)
P5	4th Moment (Kurtosis)	$\mu_4 = \sum_{i=0}^{L-1} (z_i - m)^4 p(z_i)$	(5)
P6	Relative Smoothness	$R = 1 - \frac{1}{1 + \sigma^2}$	(6)
P7	Standard Deviation	$\sigma = \sqrt{\frac{1}{K} \sum_{j=1}^{K} (z_j - m)^2}$	(7)
P8	Tamura Contrast	* Tamura, Mori, and Yamawaki, 1977	
P9	Tamura Directionality	* Tamura, Mori, and Yamawaki, 1977	
P10	Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$	(8)

Owing to the promising results obtained during our preliminary investigations (Banda and Angryk, 2009) and some earlier work (Lamb, 2008), we chose to segment our images using an eight by eight grid for our image parameter extraction and labeling. This process is illustrated in Figure 2.

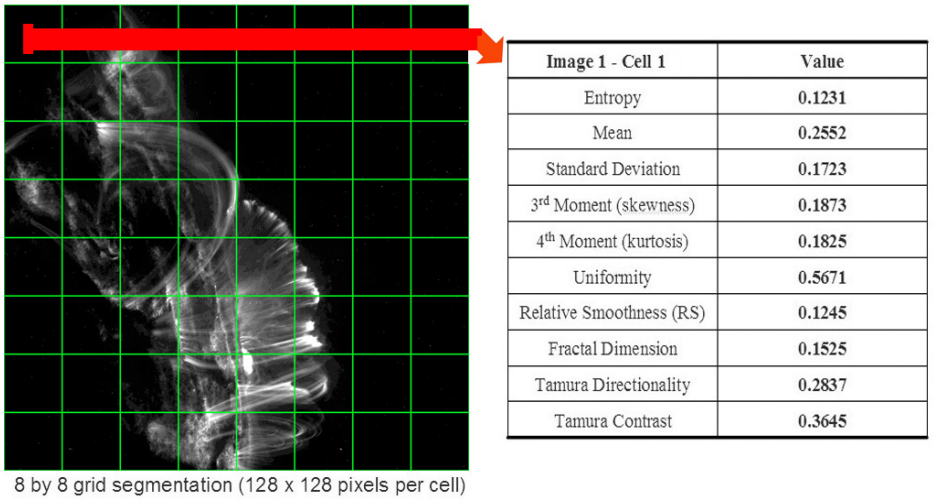


Image 1 - Cell 1	Value
Entropy	0.1231
Mean	0.2552
Standard Deviation	0.1723
3rd Moment (skewness)	0.1873
4th Moment (kurtosis)	0.1825
Uniformity	0.5671
Relative Smoothness (RS)	0.1245
Fractal Dimension	0.1525
Tamura Directionality	0.2837
Tamura Contrast	0.3645

Figure 2. Example of the conversion between solar image cells to numerical image-parameter representation.

2.2.2 Transformation from Image Parameter to Feature Vector

Once the ten image parameters have been extracted from our image dataset, we chose to treat each image parameter separately since we want to determine the usefulness and behavior of each parameter with the different dissimilarity measures individually. Each image is transformed into ten (one per image parameter from Table 2) 64-element (one cell equals one element) vectors, with each element representing the value of the each image parameter extracted from each grid cell. Each one of these vectors is treated as a histogram in the sense that each element position (1 to 64) becomes a bin and the value of each element becomes the value associated with the bin, not an actual count of them. Here the bins are representing the cell locations in the image and their values represent the statistical parameter value.

In order to use these vectors correctly when calculating some of the measures (the KLD and Jensen–Shannon divergence (JSD) measures in particular as defined in Section 2.3) we need to make sure the sum of the bins adds to one. To achieve this, we normalized every single parameter per image in the following way:

$$Normalize(A) = \frac{A_m}{\sum_{m=1}^{n} A_m} \quad for\ m = 1\ to\ n \tag{9}$$

where n = 64, since we have a total of 64 bins (dimensions in our feature vector), and A is just the parameter value at said "bin".

This allows us to scale our vectors and preserve their shape, and treat them as histograms. For "bins" equaling zero, we add a very small quantity (1×10^{-9}) in order to avoid divisions by zero on the KLD measure. Each of the ten image parameter was normalized in the same way, allowing our value ranges to be consistent and be the same for all 18 different metrics tested.

2.3 Dissimilarity Measures Used

A dissimilarity measure (in this context) is a formula that calculates how dissimilar two images are. We selected 18 dissimilarity measures for comparative analysis. Based on our literature review we find that the majority of the measures that we selected are widely used in image analysis and reported good results when applied to images in other domains (Pentland, 1984; Chaudhuri and Nirupam, 1995; Cernadas *et al.*, 2005; Wen-lun, Zhong-ke, and Jian, 2005; Holalu and Arumugam, 2006; Deselaers, Keysers, and Ney, 2008; Devendran, Hemalatha, and Amitabh, 2009; Banda and Angryk, 2010 b, c). We investigate these different measures in order to verify how well they differentiate our classes of solar phenomena as well as determine peculiarities within the classes themselves. We will address this later in our experiment section, where we present plots of dissimilarity matrices. The classical definitions of these metrics are provided in Appendix 1. Table 3 lists them and the range where they are found in the listed figures.

Table 3. Dissimilarity Measure Labels Used in Our Classification Accuracy Figures

Label	Distance	Used in Figures
D1	Euclidean	10, 12, 13, and 14
D2	Standarized Euclidean	10 and 12
D3	Mahalanobis	10 and 12
D4	City Block	10, 12, 13, and 14
D5	Chebychev	10, 12, 13, and 14
D6	Cosine	10 and 12
D7	Correlation	10 and 12
D8	Spearman	10 and 12
D9	Hausdorff	10 and 12
D10	Jensen-Shannon divergence (JSD)	10 and 12
D11	$\chi 2$	10 and 12
D12	Kullback—Leibler divergence (KLD) A-B	10 and 12
D13	Kullback—Leibler divergence (KLD) B-A	10 and 12
D14	Fractional Minkowski $p = 0.25$	13 and 14
D15	Fractional Minkowski $p = 0.50$	13 and 14
D16	Fractional Minkowski $p = 0.80$	13 and 14
D17	Fractional Minkowski $p = 0.90$	13 and 14
D18	Fractional Minkowski $p = 0.95$	13 and 14

Please note that the first eight and the last five measures are given for an m-by-n data matrix $[X]$ (in our case it contains m = 1600 images and n = 64 image parameter values), which is treated as m (1-by-n) row vectors $[x_1, x_2, ..., x_m]$. Measures 9 to 14 are presented for histograms. In order to represent our feature vector as a histogram, we treated each element of n as a bin (n = 64). For example, we convert x_s to the histogram A, the value in each bin $[A_j]$ (for $j = 1$ to n) is equal to each x_{sj} (for $j = 1$ to n).

2.4. Transformation of Dissimilarity Matrices via Multidimensional Scaling (MDS)

Since we wanted to expand our qualitative analysis of the dissimilarity matrices achieved by looking at their visualizations via the scaled image plots, we applied MDS to these dissimilarity matrices in order to verify the degree of dimensionality reduction that can be achieved with these dissimilarity measures - image parameter combinations.

MDS is a set of statistical techniques used for the exploration of dissimilarities in data, and it is popular in the field of information visualization. By using the first two most significant resulting components, we get the standard 2D MDS plots; if we use the first three most significant components we get the 3D MDS plots. These "components" are the sorted eigen-values whose relative magnitudes indicate how many dimensions we can safely use. MDS is also commonly used as

a method for dimensionality reduction for large dissimilarity matrices (Beatty and Manjunath, 1997; Rubner, Guibas, and Tomasi, 1997; Borg and Groenen, 2005; Datta, Li, and Wang, 2005;). We use the classical MDS approach, since we have input matrices giving dissimilarities between pairs of items (produced by our dissimilarity measures). This process will output a coordinate matrix whose configuration minimizes a loss function called strain.

With the resulting MDS matrices, we have a new re-arranged dimensional space, based on the dissimilarity matrices of the original data (similar to Principal Component Analysis (PCA) or Singular Value Decomposition (SVD)). However, one of the main issues behind MDS is that does not provide an explicit mapping function governing the relationship between patterns in the input space and in the projected space (Naud, 2001). This issue significantly limits the popularity of MDS as a dimensionality-reduction technique, since you need the full dissimilarity matrix of the total data to generate the new dimensional spaces.

2.5. Classification Algorithms as a Tool for Quantitative Evaluation

In general terms, a classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations. In our context, a classification algorithm will build a model based on labeled-training data that will help with the proper labeling of new data (test set). In this article, we selected the Naïve Bayes Classifier and Support Vector Machines (SVM) with a linear kernel function as our linear classifiers, and C4.5 as a decision tree classifier. Linear classifiers create groupings of items that have similar feature values by making classification decisions based on the value of a linear combination of the features. C4.5 uses an entropy-based information gain measure to split samples into classes.

In terms of how to measure these classification experiments, we selected the traditional classification accuracy measure to determine their effectiveness.

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}} \tag{10}$$

The accuracy of classification measure derives from a **confusion matrix.** In this matrix, each column represents the instances in a classified class, while each row represents the instances in the actual class as shown in Figure 3.

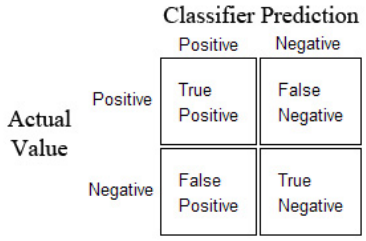


Figure 3. Confusion matrix example

Note that we use these classifiers in order to present a comparative and quantitative analysis of our experiments, and we are not trying to find the best classification results or adjust the classifiers to perform at their best. We are trying to determine two things: i) the winning combination of dissimilarity measure / image parameter, and ii) an estimate of how many components of our new artificial MDS data space can be omitted without a significant decrease in the classification accuracy.

3. Experimental Evaluation and Results

In this section, we will provide the details of our experimental evaluation and the results we obtained.

3.1. Preliminaries

All experiments were performed using Matlab© R2009b. For the exponential curve fitting we used the Ezyfit Toolbox (www.mathworks.com/matlabcentral/fileexchange/10176).

The classification experiments were performed using WEKA 3.6.1 (Hall *et al.* 2009). These programs were run on a PC with an AMD Athlon II X4 2.60 GHz Quad Core processor with 8 GB of RAM and Windows XP 64-bit Edition.

After all our data was normalized, we calculated the pairwise distance between the histograms using Matlab's pdist function. As this function is highly optimized for performance, the computation time for our first eight (and last five) measures was less than a few minutes. The Haussdorff, KLD, JSD and χ^2 distances were implemented from scratch and yielded a higher computational expense due to the nature of the algorithms.

In total we produced 180 dissimilarity matrices (18 measures, counting KLD A-B and B-A separately, times a total of ten different image parameters). All these dissimilarity matrices are symmetric, real and positive valued, and have zero-valued diagonals. Thus they fit the classical MDS requirements.

3.2. Dissimilarity Matrix Calculations

Figure 4 presents these dissimilarity matrices that help us to qualitatively identify which image parameters and measures provide informative differentiation for our images between the eight different classes of our dataset. In this article, we show three of the most interesting parameter-measure combinations generated (good and bad). We refer readers who want to learn about specific combinations to our presentation online at (www.jmbanda.com/SDOJournal2011/), where all 180 combinations are presented. Here the classes of our benchmark are separated on the axes. Blue means low dissimilarity, and red means high dissimilarity. Please note that each dissimilarity matrix has been normalized using min-max normalization in order to have all figures scale to the same range, allowing us to compare all different dissimilarity metrics in the same context and all figures to be consistent.

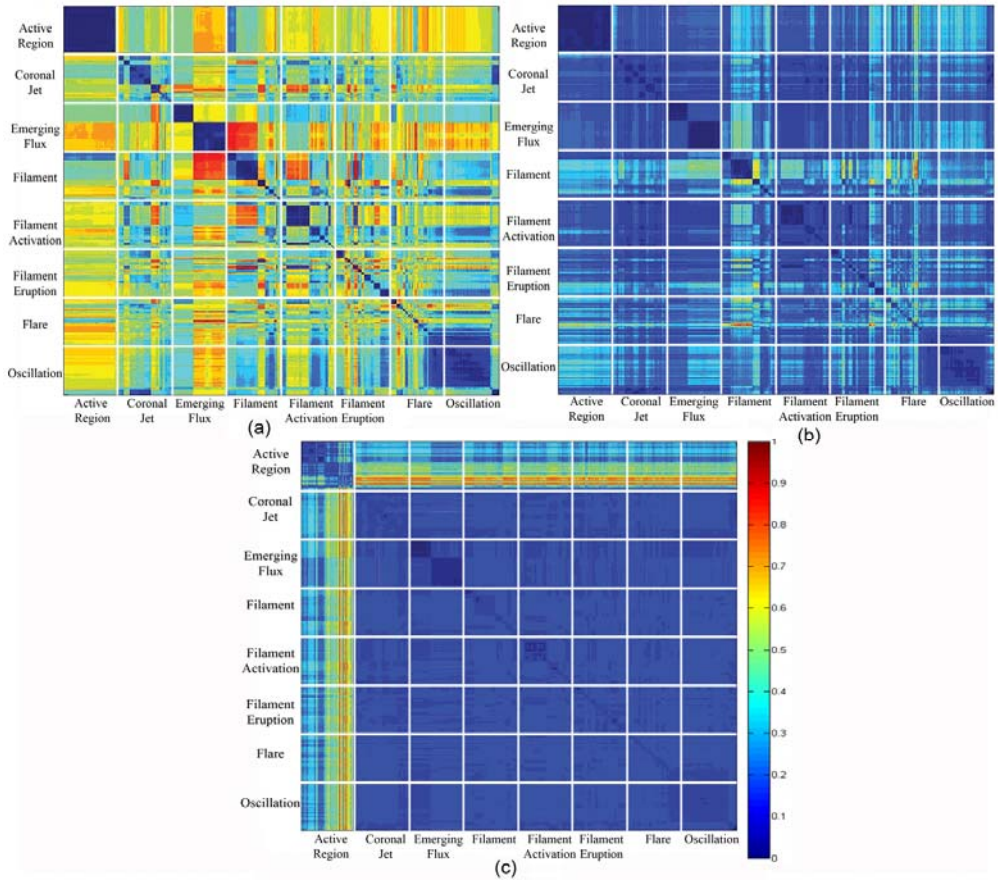


Figure 4. Scaled dissimilarity matrix for (a) Correlation (D7) measure with image parameter mean (P3), (b) *JSD* (D10) measure with image parameter mean (P3), (c) Chebychev (D5) measure with image parameter relative smoothness (P6). The color bar next to (c) represent the dissimilarity ranges for all Figures with these kinds of plots.

As we can see from figure 4(a), this combination of similarity measure D7 (correlation) and image parameter P3 (mean) produces an example of poor separation between all classes since we have very diverse coloring throughout the figure. Figure 4(b) shows that the D10 (JSD) measure produces an entirely different dissimilarity matrix for the same image parameter P3 (mean), which highlights different dissimilarities than the correlation measure (Figure 4(a)). Figure 4(c) is a clear example of a combination of a dissimilarity measure D5 (Chebychev) and an image parameter P6 (relative smoothness) that highlights dissimilarities within only one class of the benchmark, but recognizes that everything else is highly similar for the rest of the classes. In these plots we are looking for clear blocks that separate one class from the others, as found in Figure 4(c). This validates our idea of testing every pair of dissimilarity measures and image parameters individually, since there are combinations that will allow us to notice different relationships between the classes of solar images.

3.3. Transformation of Dissimilarity Matrices via Multidimensional Scaling (MDS)

After generating 180 dissimilarity matrices, we performed classical multidimensional scaling using Matlab's cmdscale function. MDS has been widely

utilized in image-retrieval research to reduce dimensionality (Beatty and Manjunath, 1997; Rubner, Guibas, and Tomasi, 1997), and to aid in the visual depiction of image dissimilarity in a convenient two and three dimensional plot (Borg and Groenen, 2005). However, these articles present results on a considerably smaller number of images and use a considerably smaller number of dimensions.

The most commonly used MDS plots involve using the first two or three components of the calculated coordinate matrix. In Figure 5 we show both 2D and 3D plots of these components.

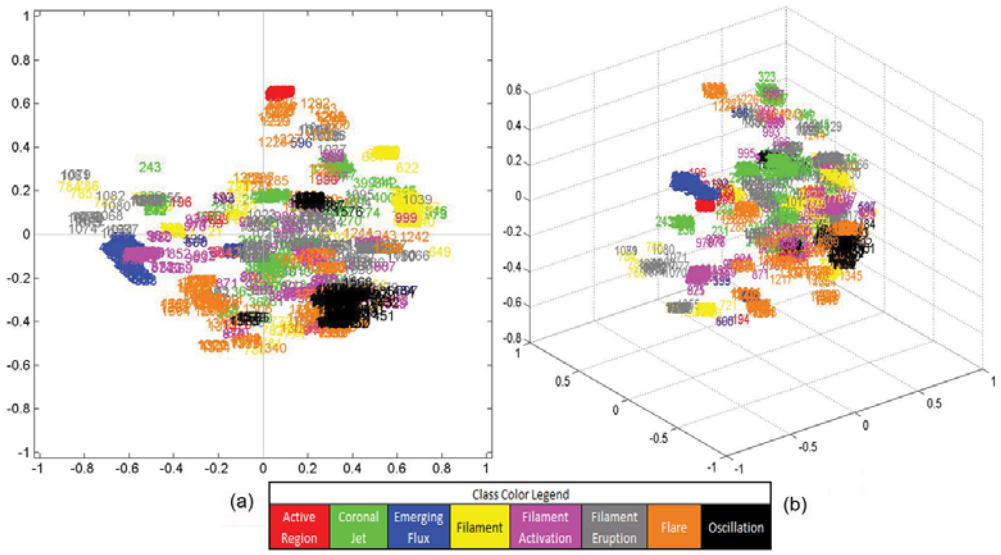


Figure 5. MDS map for the correlation measure D7 with image parameter mean (P3) (a) top 2 and (b) top 3 components. Each image in each class is represented by a number and each class is described in different colors as indicated on the legend.

As we expected, in Figure 5 we cannot easily identify a clear separation between our eight different classes on two or three dimensions of the resulting MDS dimensional space. However, selecting the correct combination might yield interesting 3D maps for certain dissimilarity measure and image parameter pairs. In Figure 6 we show the 3D MDS component plots for the dissimilarity matrices of Figure 4. Here we show that, while for Figure 6 (a) and (b) the maps do not really highlight any clusters, in (c) we have a clear cluster for class *Active Region* (color red). This is exactly what is indicated in Figure 4 (c). These 3D MDS components plots sometimes show very clear clusterings, and other times they can be used to aid in the interpretation of the scaled image plots of the dissimilarity matrices. (We will talk about this when we discuss the fractional Minkowski metrics results)

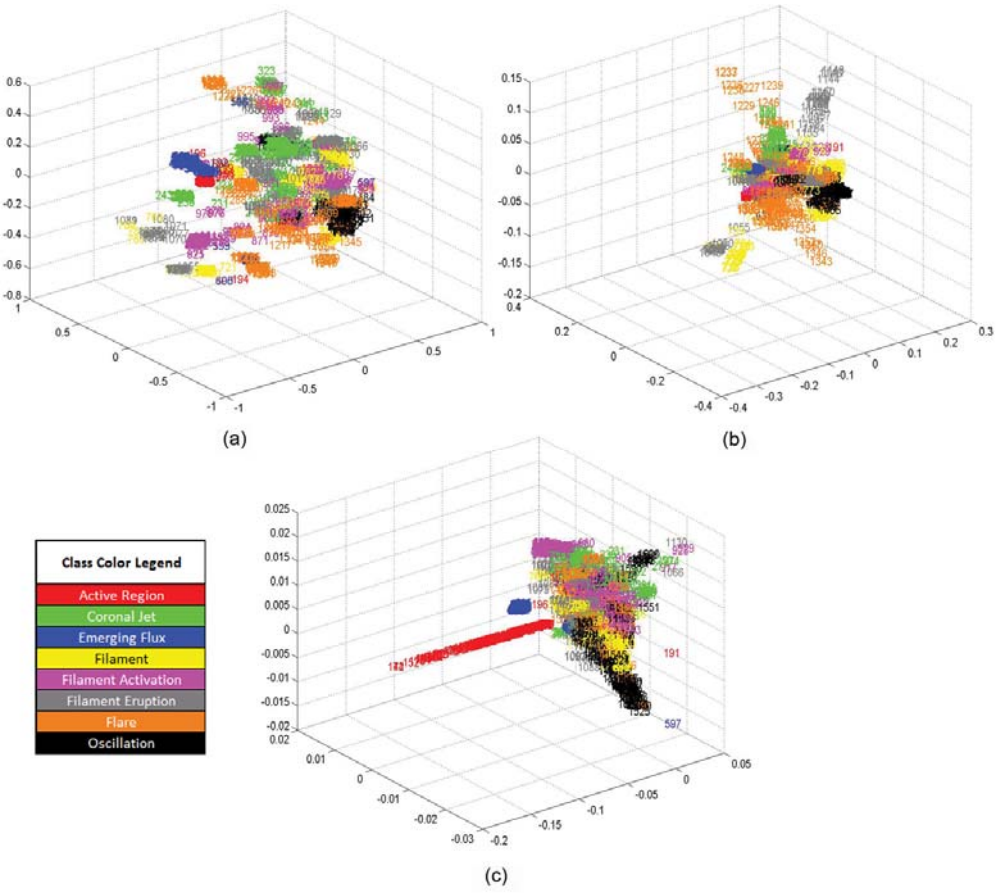


Figure 6. 3D MDS components plot for (a) Correlation (D7) measure with image parameter mean (P3), (b) JSD (D10) measure with image parameter mean (P3), (c) Chebychev (D5) measure with image parameter relative smoothness (P6). Each image in each class is represented by a number and each class is described in different colors as indicated on the legend.

3.4. Fractional Minkowski Metrics

We investigate the behavior of Minkowski-based fractional metrics on our domain specific dataset, because they have provided very interesting results in other domains (Aggarwal, Hinneburg, and Keim, 2001; Francois, Wertz, and Verleysen, 2005), and are mathematically proven to be better than other more traditional non-fractional Minkowski metrics (Aggarwal, Hinneburg, and Keim, 2001). In this section, using Figures 7 and 8, we present a comparison between the scaled image plot and the 3D MDS components plot for two different fractional metrics of the same image parameter. In these two figures we illustrate how the fractional metrics improve in 'performance' when p gets closer to one (City Block dissimilarity measure (D4)).

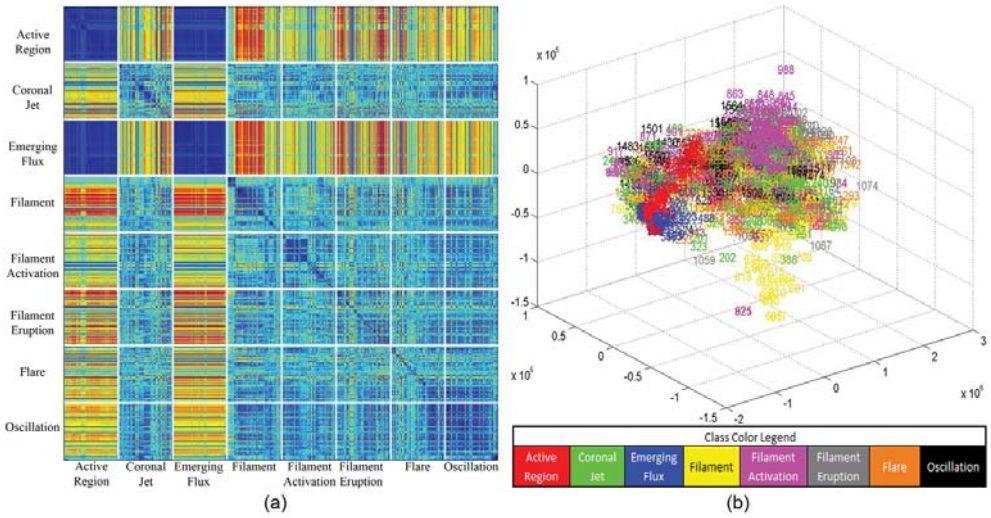


Figure 7. Fractional Minkowski metric with p = 0.25 (D14) and the fractal dimension image parameter (P2): (a) Scaled Image Plot, (b) 3D MDS components plot. The color legend only applies to (b).

From Figure 7, we can observe a relatively clean separation of two classes (Active Region and Emerging Flux), when it comes to the scaled image plot of the dissimilarity matrix (a). However, in terms of the 3D MDS components plot (b) the clusters corresponding to Active Region (red color) and Emerging Flux (blue color) seem somewhat fuzzy and mixed in with the rest of the classes. According to our literature review (Aggarwal, Hinneburg, and Keim, 2001; Francois, Wertz, and Verleysen, 2005) these fractional metrics tend to improve when p approaches one. In Figure 8 we try to verify these claims for our own domain-specific dataset.

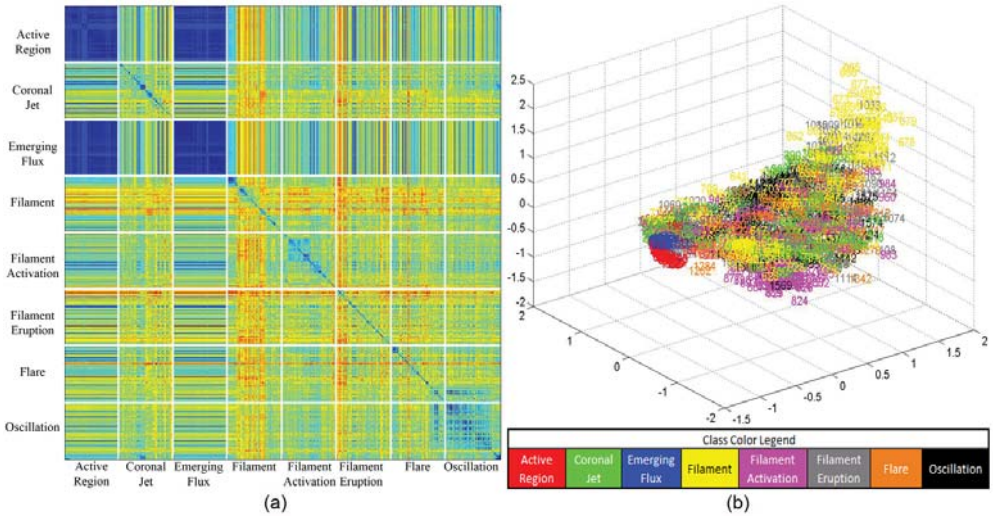


Figure 8. Fractional Minkowski metric with p = 0.95 (D18) and the fractal dimension image parameter (P2): (a) Scaled Image Plot, (b) 3D MDS components plot. The color legend only applies to (b).

As the literature mentions (Aggarwal, Hinneburg, and Keim, 2001; Francois, Wertz, and Verleysen, 2005), once p is approaches one we see that on both the scaled image plot of the dissimilarity matrix (Figure 8 (a)) and the 3D MDS components plot (Figure 8 (b)), the classes Active Region and Emerging Flux are more separable and hence have better clustering. It is also very interesting that the rest of the classes tend to get "dragged away" from these two classes, making the

Active Region and Emerging Flux classes even more distinguishable. Note however, this increases the separation of the other classes. This only shows that each combination of dissimilarity measure and image parameter is unique, and that their behaviors need to be analyzed individually to gain the most information.

3.5. Component Thresholding

Based on the magnitude of each of the resulting MDS components, we decided to use exponential curve fitting in order to apply a threshold to the optimal number of components needed to reduce dimensionality and still retain valuable components to produce good classification results. For comparative purposes, we also investigated a far simpler approach of selecting only the top ten components and discarding the rest. These two approaches allow us to verify how much a few (or many) extra components will increase or decrease the accuracy of our classification results, or in other words how many components we need in order to maintain a good representation of our data in the new dimensional space, and how many components we can discard in order to optimize our querying responsiveness. Note that this MDS dimensionality-reduction analysis is of exploratory nature, and it is only used for estimation of potential dimensionality reduction. We present a more comprehensive and extensive analysis using better fitted methods in Banda, Angryk, and Martens, 2012.

In order to determine the aforementioned number of components, we plotted the magnitude of each component. Since the MDS matrix output is ordered by importance, the magnitudes should be decreasing as the number of components increases. After empirical analysis of the magnitude of the resulting MDS components, we observed that after ten components the decrease of their magnitudes moderates (in most of the cases), so therefore we initially decided to take a somewhat naïve approach and applied a threshold of ten components per similarity measure/image parameter combination.

In our second approach, we utilized exponential curve fitting (Shepard, 1980) to find a function that models this behavior. Our intent was to locate a threshold for the number of necessary components. We utilize a 135-degree angle of the tangent line to this function to determine the threshold and discard the components whose magnitudes are not providing significant improvement over the previous ones. This tangent line is indicated as the red line in Figure 9; the red dot indicates the intersection point, which equals the number of components that we will use.

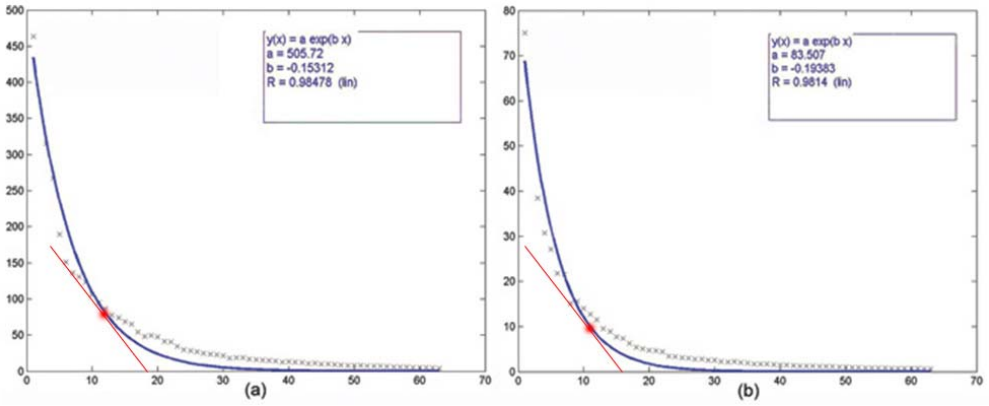


Figure 9. The set of sorted (descending) eigenvalues whose relative magnitudes indicate how many components (dimensions) one can safely use. Each eigenvalue is indicated by the *y*-axis, and the number of components is indicated by the *x*-axis (from 1 to 63). Top is the exponential curve fitting for: (a) correlation measure (D7) with image parameter mean (P3), (b) JSD measure (D10)

with image parameter mean (P3). The red tangent line is used to select the number of components to analyze at the intersecting point (red dot).

As can be seen from Figure 9, the magnitudes of the components (*y*-axis) decrease up to a certain point, and after this point the change is minimal and thus not too important for the new dimensional space.

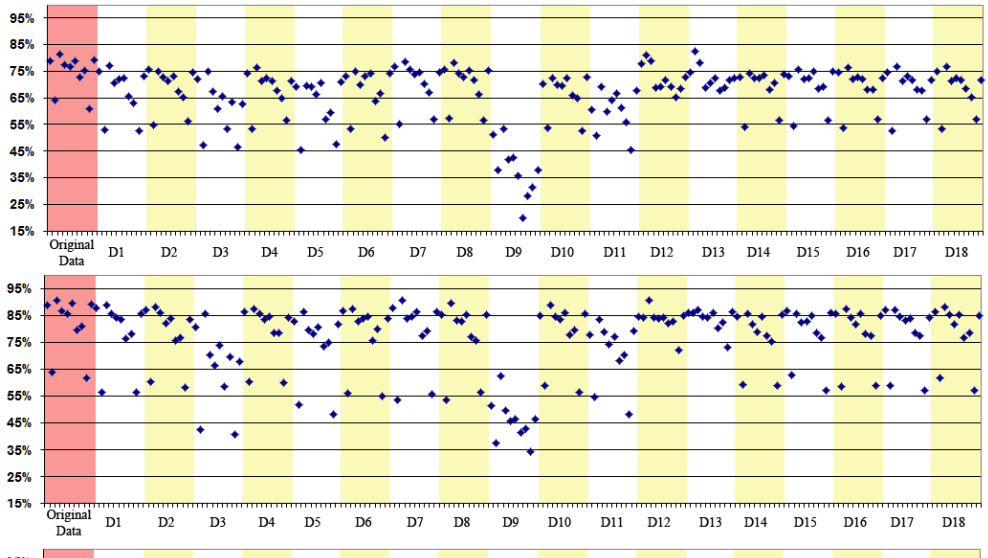
Based on these curve-fitting results and the threshold output, we determined a specific number of components per combination of dissimilarity-measure—image-parameter. We can now determine how well this reduced dimensionality performs in our classification tasks in Section 3.6.

3.6. Quantitative Evaluation via Comparative Analysis of Classifiers

We have described how we applied the dissimilarity measures to our image parameters and produced dissimilarity matrices. We also analyzed how MDS transformed these dissimilarity matrices into a different dimensional space, one that will require, hopefully, fewer dimensions in order to distinguish different types of solar phenomena. We now describe the classification experiments we performed on the naïve ten-component threshold and the tangent thresholded components compared to our original data. All classification experiments used tenfold cross-validation.

We ran a total of 180 different datasets through the three classifiers described in Section 2.5. In the following figures we present the overall results of these classification experiments and offer a more detailed explanation of the most interesting results.

Figure 10 shows the classification accuracy of our selected classifiers on our ten component per dissimilarity-measure—image-parameter combination. The first ten columns indicate our original normalized dataset values with no measure or dimensionality reduction applied to them. The rest of the columns indicate our dissimilarity-measure—image-parameter combinations in the. Each tick in each group represents one parameter from Table 2. Each group contains ten parameters and then it moves on to the next dissimilarity measure as indicated in the figure.

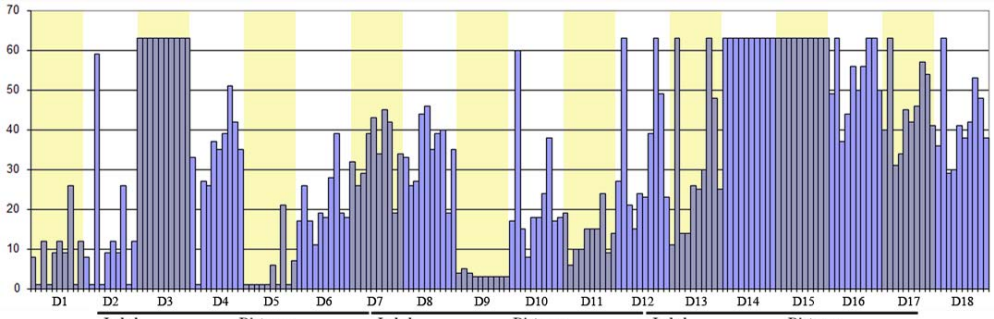


Label	Distance	Label	Distance	Label	Distance
D1	Euclidean	D7	Correlation	D13	Kullback-Leibler divergence (KLD) B-A
D2	Standarized Euclidean	D8	Spearman	D14	Fractional Minkowski *p=0.25*
D3	Mahalanobis	D9	Hausdorff	D15	Fractional Minkowski *p=0.50*
D4	City Block	D10	Jensen-Shannon divergence (JSD)	D16	Fractional Minkowski *p=0.80*
D5	Chebychev	D11	$\chi 2$	D17	Fractional Minkowski *p=0.90*
D6	Cosine	D12	Kullback-Leibler divergence (KLD) A-B	D18	Fractional Minkowski *p=0.95*

Figure 10. Percentage of correctly classified instances for the ten-component threshold: (top) for the Naïve Bayes classifier, (middle) for the decision-tree (C4.5) classifier, and (bottom) for the SVM classifier, for the original data and our 180 experiments (D1 to D18).

As shown in Figures 10(a) and 10(b), our ten-components-only approach produces very similar classification results to our original data for most combinations of measure and image parameters. We also notice that the worst performing dissimilarity-measure—image-parameter combination is D9, corresponding to the Hausdorff dissimilarity measure.

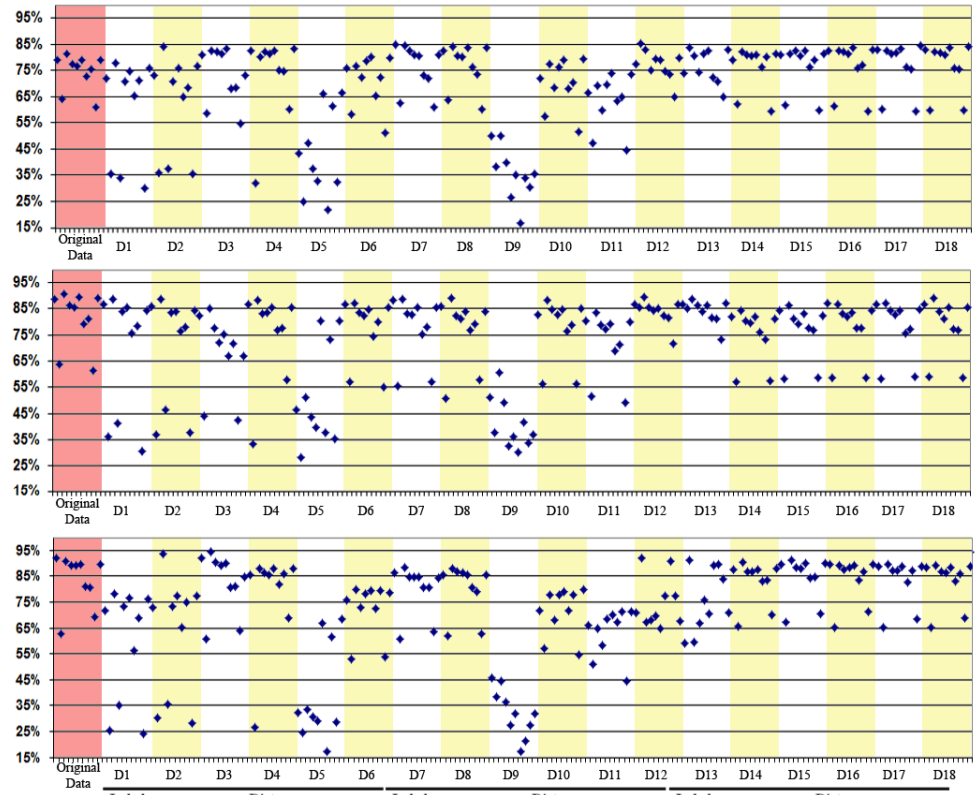
In figure 11 we show the resulting number of components to be used based on the tangent thresholding. The columns represent the 180 different image parameter/measure combinations (with the omission of the first ten, which are the original dataset). In this figure, a high number of components indicates that the components do not seem to decrease steadily in magnitude.



Label	Distance	Label	Distance	Label	Distance
D1	Euclidean	D7	Correlation	D13	Kullback-Leibler divergence (KLD) B-A
D2	Standarized Euclidean	D8	Spearman	D14	Fractional Minkowski *p=0.25*
D3	Mahalanobis	D9	Hausdorff	D15	Fractional Minkowski *p=0.50*
D4	City Block	D10	Jensen-Shannon divergence (JSD)	D16	Fractional Minkowski *p=0.80*
D5	Chebychev	D11	$\chi 2$	D17	Fractional Minkowski *p=0.90*
D6	Cosine	D12	Kullback-Leibler divergence (KLD) A-B	D18	Fractional Minkowski *p=0.95*

Figure 11. Number of components to use indicated by the tangent-thresholding method for each distance measure-image parameter combination.

In Figure 12 we show the tangent-thresholded classification results. The number of components selected varies between 1 and 63 depending on the combination of measure/image parameter. For direct comparison the experiments were ordered the same way as in Figure 10.



Label	Distance	Label	Distance	Label	Distance
D1	Euclidean	D7	Correlation	D13	Kullback-Leibler divergence (KLD) B-A
D2	Standarized Euclidean	D8	Spearman	D14	Fractional Minkowski *p=0.25*
D3	Mahalanobis	D9	Hausdorff	D15	Fractional Minkowski *p=0.50*
D4	City Block	D10	Jensen-Shannon divergence (JSD)	D16	Fractional Minkowski *p=0.80*
D5	Chebychev	D11	$\chi 2$	D17	Fractional Minkowski *p=0.90*
D6	Cosine	D12	Kullback-Leibler divergence (KLD) A-B	D18	Fractional Minkowski *p=0.95*

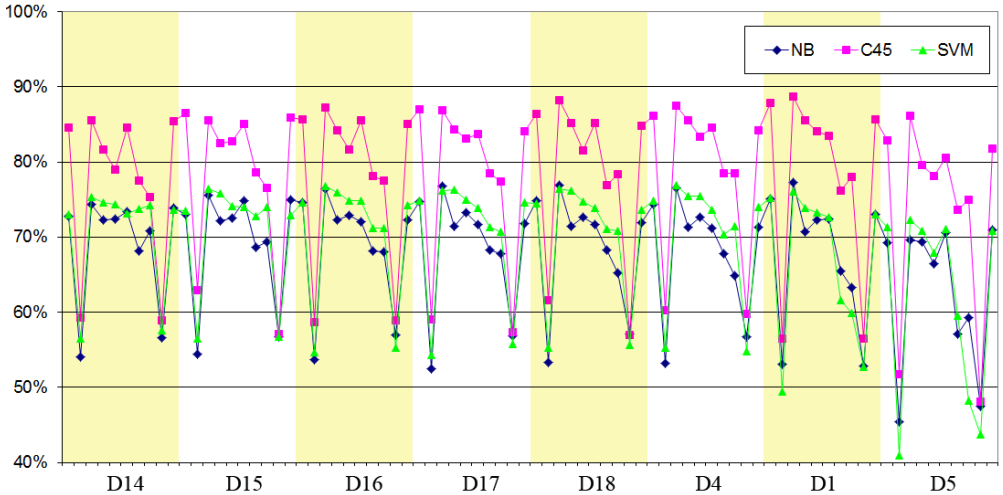
Figure 12. Percentage of correctly classified instances for the tangent-based component threshold: (top) for the Naïve Bayes classifier, (middle) for the decision-tree (C4.5) classifier, and (bottom) for the SVM classifier, for the original data and our 180 experiments (D1 to D18).

In these thresholded components classification results we get very similar results compared to classification results that used only ten components for NB and C4.5, and in some cases we get considerable drops like for the Chebychev measure (D5). This is due to the fact that the tangent based thresholding selects fewer than ten components per combination of measure/image parameter and in some instances even only one component (see Figure 11). An interesting thing to notice is that the overall classification percentages increase consistently for the KLD A-B and KLD B-A (D12 and D13) combinations, although this might be due to the fact that the tangent thresholding selected 63 components for several of the image parameter combinations. We also observe that providing large numbers of components to greedy classifiers (NB and C4.5), making the locally optimal choice at each stage, does not help them improve much, whereas SVM take clear advantage of more data.

In the results for both the tangent thresholding and the ten-component limiting, we observe that with only ten components we can achieve good accuracy results (around 80 to 90 %) for the selected classifiers. This translates to an estimated average of 70 % dimensionality reduction from our original number of dimensions. We can also see which image parameters perform the best with which measures, which was one of our objectives with this research.

3. 7. Quantitative Evaluation via Comparative Analysis of Classifiers for the Fractional Minkowski Metrics

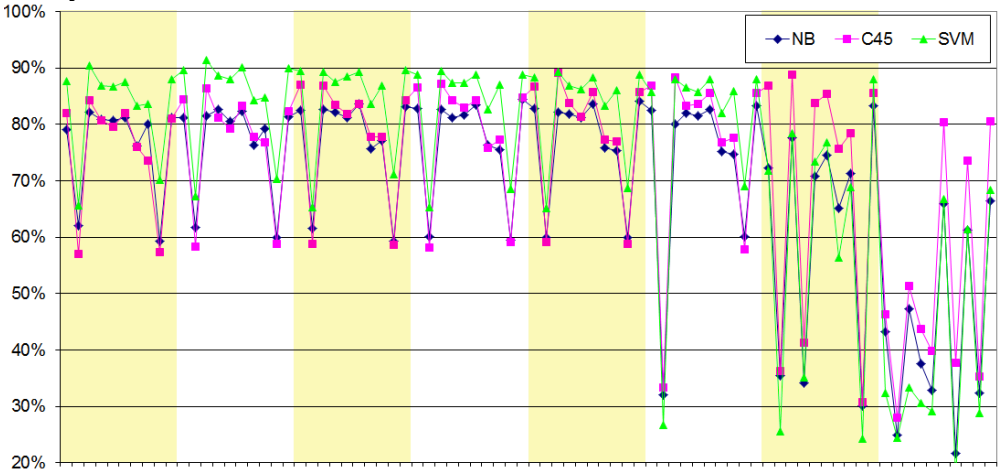
In Figures 13 and 14, we take an in-depth look at the classification results for the Fractional Minkowski dissimilarity measures (D14 to D18) paired with our other three dissimilarity measures based on the Minkowski metric. These dissimilarity measures are: City Block distance ($p = 1$) (D4), Euclidean distance ($p = 2$) (D1), and Chebychev measure ($p = \infty$) (D5).



Label	Distance	Label	Distance
D1	Euclidean	D15	Fractional Minkowski $p=0.50$
D4	City Block	D16	Fractional Minkowski $p=0.80$
D5	Chebychev	D17	Fractional Minkowski $p=0.90$
D14	Fractional Minkowski $p=0.25$	D18	Fractional Minkowski $p=0.95$

Figure 13. Percentage of classification accuracy for the ten-component Minkowski based fractional metrics with $p = 0.25$ to $p = \infty$ (D14 — 18, D4, D1, D5), ordered by image parameter.

As we can see in Figures 9 and 11 (D14-D18) and in Figure 12, the fractional metrics tend to perform in a very stable manner, dropping for the same image parameters (Fractal Dimension (P2) and Tamura Directionality (P9)) for all of the different p values we investigated. An interesting thing in this figure is that as p gets closer to one, the experiments do not show a clear tendency of increasing classification accuracy. Also, while we have different results for intra-class separation for some class's behavior, as we show in Section 3.4, the inter-class separation of the remaining classes seems to balance out the classification results. We also have bigger drops as p approaches infinity for the Chebychev dissimilarity measure (D5), indicating that Minkowski metrics with a p closer to 2, perform better (and more stably) than this dissimilarity measure. A final observation from Figure 12 is that with the ten-component threshold the tree-based classifier (C4.5) achieves the best classification results, while SVM seems to stay almost 10 % behind.



Label	Distance	Label	Distance
D1	Euclidean	D15	Fractional Minkowski $p=0.50$
D4	City Block	D16	Fractional Minkowski $p=0.80$
D5	Chebychev	D17	Fractional Minkowski $p=0.90$
D14	Fractional Minkowski $p=0.25$	D18	Fractional Minkowski $p=0.95$

Figure 14. Percentage of classification accuracy for the Tangent thresholded Minkowski based fractional metrics with $p= 0.25$ to $p=\infty$ (D14 — 18, D4, D1, D5), ordered by image parameter.

In Figure 14, we see two very interesting differences between the tangent threshold and our ten-component threshold. The first is that SVM takes a definite lead in terms of classification, with accuracy approaching 90 % for the majority of dissimilarity measure/image parameters pairs and reaching our overall best result of 94.4 % for the mahalanobis dissimilarity measure (D3) combined with the mean image parameter (P2). The tree-based classifier C4.5 keeps almost the same classification accuracy as with ten-components, causing us to believe that high dimensionality does not benefit this classification model. The improvement of the SVM accuracy shown in Figure 14, is due to the fact that we selected a considerably larger amount of components with the tangent threshold, between 46 and 63, with the majority being 63 (see Figure 9). This greatly improves the performance of the SVM classifier in the majority of our experiments. The selection of this large number of components is due to the fact that the decrease in magnitude of the components is very small, making curve fitting ineffective. The second thing to notice is that our classification accuracy drops considerably for p

values greater and equal to one, over 30 — 40 % for some image parameters, and for the majority of the Chebychev measure (D5) results. This could be partly due to the fact that the number of components selected is less than 30 for most instances (see Figure 14, experiments for $p = 1$ (D4), $p = 2$ (D1), and for $p = \infty$ (D5)), as well as the actual dissimilitude found for the combination of measure/image parameter in question. In Table 4 we present the top ten classification results for both the tangent thresholded and the ten components limited datasets. This will help us to quantitatively evaluate the differences between these two methods.

Table 4. Top ten classification accuracy results for ten component thresholded dimensionality reduction experiments. DM indicates dissimilarity measure and IP indicates image parameter.

NB		C45		SVM	
DM-KLD B-A-IP-FracDim	82.44	DM-correlation-IP-Mean	90.63	DM-spearman-IP-Mean	81.13
DM-KLD A-B-IP-FracDim	81.25	DM-KLD A-B-IP-Mean	90.63	DM-correlation-IP-Mean	80.19
DM-KLD A-B-IP-Mean	78.88	DM-spearman-IP-Mean	89.63	DM-correlation-IP-Entropy	77.88
DM-correlation-IP-Mean	78.75	DM-Euclidean-IP-Mean	88.69	DM-spearman-IP-RelSm	77.44
DM-spearman-IP-Mean	78.19	DM-JSD-IP-Mean	88.63	DM-spearman-IP-Uniformity	77.44
DM-KLD B-A-IP-Mean	78.06	DM-StandEuclidean-IP-Mean	88.25	DM-correlation-IP-RelSm	77.13
DM-KLD A-B-IP-Entropy	77.81	DM-Min p=0.95-IP-Mean	88.25	DM-correlation-IP-Moment3	77.00
DM-Euclidean-IP-Mean	77.25	DM-Euclidean-IP-Entropy	87.88	DM-correlation-IP-Uniformity	77.00
DM-correlation-IP-Entropy	76.94	DM-correlation-IP-Entropy	87.69	DM-cityblock-IP-Mean	76.88
DM-Min p=0.95-IP-Mean	76.88	DM-cityblock-IP-Mean	87.50	DM-spearman-IP-Entropy	76.88

Table 5. Top ten classification accuracy results for tangent thresholded dimensionality reduction experiments. DM indicates dissimilarity measure, IP indicates image parameter, and C= indicates number of components.

NB		C45		SVM	
C=63-DM-KLD A-B-IP-FracDim	85.50	C=21-DM-KLD A-B-IP-Mean	89.38	C=63-DM-mahalanobis-IP-Mean	94.44
C=32-DM-correlation-IP-Entropy	85.06	C=27-DM-spearman-IP-Mean	89.19	C=59-DM-StandEuclidean-IP-Mean	93.88
C=29-DM-correlation-IP-Mean	84.69	C=12-DM-euclidean-IP-Mean	88.81	C=63-DM-mahalanobis-IP-Entropy	92.13
C=59-DM-STDEuclidean-IP-Mean	84.31	C=59-DM-STDEuclidean-IP-Mean	88.75	C=63-DM-KLD A-B-IP-FracDim	92.13
C=27-DM-spearman-IP-Mean	84.19	C=14-DM-KLD B-A-IP-Mean	88.75	C=63-DM-KLD B-A-IP-FracDim	91.38
C=63-DM-KLD B-A-IP-FracDim	83.94	C=29-DM-correlation-IP-Mean	88.63	C=63-DM-KLD A-B-IP-TamCont	91.00
C=35-DM-spearman-IP-RelSm	83.75	C=15-DM-JSD-IP-Mean	88.31	C=63-DM-mahalanobis-IP-Moment3	90.56
C=35-DM-spearman-IP-Uniformity	83.75	C=27-DM-cityblock-IP-Mean	88.25	C=63-DM-mahalanobis-IP-RelSm	90.06
C=63-DM-mahalanobis-IP-RelSm	83.25	C=32-DM-correlation-IP-Entropy	88.25	C=63-DM-KLD B-A-IP-TamCont	89.50
C=35-DM-cityblock-IP-Uniformity	83.25	C=25-DM-KLD B-A-IP-Uniformity	87.31	C=63-DM-mahalanobis-IP-Moment4	89.38

In Table 4 (the ten-component threshold) we have very different measures performing the best for the different classifiers. In terms of our image parameters, Mean manages to appear in our top ten almost 50 % of the times, next to fractal dimension (P2), and entropy (P1). It is interesting to note that the KLD measure (D12 - 13) seems to perform very well for our Naïve Bayes classifier, filling five out of 10 spots in this table. KLD also takes the first two spots when combined with the fractal dimension parameter (P2), and both directions of this KLD measure (A-B and B-A, D12 and D13) seem to appear close to each other, indicating that we might be able to only use one of the directions and considerably reduce its computational expense. We noticed that for the remaining two classifiers (C4.5 and SVM) we have different combinations of the mean parameter (P3) and very different types of measures (correlation (D7), KLD (D12-13) and Spearman (D8)) taking the first few places. However, the most interesting result for us, was that the fractional metrics only appear twice in this table, showing that

they are not very good for our dataset. They are also easily beaten by the euclidean distance (P1) on both occasions, showing a different trend than in the artificial dataset comparison presented in by Aggarwal, Hinneburg, and Keim, (2001) and Francois, Wertz, and Verleysen, (2005) in terms of classification accuracy.

The tangent thresholded classification results in Table 5, show different trends in the top ten classification results. The combinations presented for the Naïve Bayes classification results only beat the ten-component ones by 3 - 4 % and have almost three times as many components (from 29 to 63), showing that the ten-component threshold performs well when it comes to this classifier. We also have surprising results for the C4.5 classifier. Here, for the top results, we actually have a drop in accuracy of 1 % when using three times as many components (first two), showing the behavior we mentioned about trees keeping similar classification results for both thresholding methods (but with a considerable increase in the number of data points). The combination of dissimilarity-measure—image-parameter seems to hold for these two classifiers when it comes to achieving the highest accuracy. One thing to note is that none of the fractional metrics (D14 — 18) make any difference when it comes to this thresholding method. This reinforces the claim that they are very stable (hence their classification accuracy is not increasing for a higher number of components selected by the tangent thresholding method).

4. Conclusions

With the ambitious tasks of analyzing all of the combinations between image parameters and dissimilarity measures completed, we managed to create a solid foundation that will allow us to determine what works best for quick and accurate recognition of different solar phenomenon. The results of these experiments also show that we can reduce our dimensionality considerably and still achieve good classification results.

Some dissimilarity measures, such as correlation (D7), euclidean (D1), KLD (D12 - 13), and JSD (D10), allow us to find the dissimilarities between the images in our dataset and provided different levels of relevance for different image parameters. As every applied researcher knows, not everything always works, and with this research we can actually distinguish what works well and when in terms of solar images. The fractional dissimilarity measures perform very well for some datasets according to the literature (Aggarwal, Hinneburg, and Keim, 2001; Francois, Wertz, and Verleysen, 2005), but in our specific domain they do not seem to significantly make any significant impact to enable us to improve our results over the traditional Minkowski-based measures (city block (D4), euclidean (D1)).

While not all dissimilarity measures performed equally well, we now know which ones to omit due to their computational expense for future experiments (i.e. Hausdorff measure (D9)).

In terms of dimensionality reduction, we managed to achieve very similar classification results to those obtained with the original data. A similar machine learning approach to classify by individual phenomenon can benefit from our approach on how to select the number of components and implement it in order to speed up query retrieval times.

With the massive number of experiments performed, we lack the proper space in this medium to display all the results we produced. All the dissimilarity matrices,

MDS maps, exponential curve fitting plots, and classification results are available (www.jmbanda.com/SDOJournal2011/) for researchers interested in all these results. We also included all Matlab and WEKA files produced in order for researchers to replicate these results easily.

5. Future Work

We are currently working with dimensionality-reduction methods other than MDS, such as Principal Component Analysis and Singular Value Decomposition. These two methods have the advantage of producing mapping functions in order to transform new data into the artificial dimensional space created by them. This will enable us to use a particular training dataset and a new test dataset in order to create more accurate classification predictions.

As mentioned above, all of the classifiers used in this article were created using their default WEKA settings. The classification results are for comparative purposes and in no way they reflect the results that can be obtained after fine tuning the settings of these classifiers. We are currently working on this, and we expect to publish soon results of fine-tuned classifiers. We also expect to increase the number of classifiers used to have a more comprehensive evaluation of them.

Lastly, we continue working towards the goal of creating a fully working CBIR system for the SDO mission, and with this work as well as our previous articles, we are getting closer to this goal.

Acknowledgements: This work was supported in part by the NASA Grant Award No. 08-SDOSC08-0008, funded from NNH08ZDA001N-SDOSC: Solar Dynamics Observatory Science Center solicitation. We would also like to thank our internal reviewers Michael Schuh and Richard McAllister.

Appendix 1

Classical definitions of dissimilarity measures used in this work:

1) ***Euclidean distance:*** As found in Yang and Trewn (2004) it is defined as the distance between two points given by the Pythagorean Theorem:

$$D1 = \sqrt{(x_s - x_t)(x_s - x_t)'} \tag{10}$$

2) ***Standardized Euclidean distance:*** As found in Yang and Trewn (2004) it is defined as the Euclidean distance calculated on standardized data by the standard deviations:

$$D2 = \sqrt{(x_s - x_t)V^{-1}(x_s - x_t)'} \tag{11}$$

Where V is the n-by-n diagonal matrix whose j^{th} diagonal element is $S(j)^2$, where S is the vector of standard deviations.

3) ***Mahalanobis distance:*** As found in Yang and Trewn (2004) this is the Euclidean distance with normalization based on a covariance matrix making it scale-invariant:

$$D3 = \sqrt{(x_s - x_t)C^{-1}(x_s - x_t)'} \tag{12}$$

Where C is the covariance matrix.

4) ***City block distance:*** As found in Yang and Trewn (2004) it represents the distance between points in a grid by examining the absolute differences between coordinates of a pair of objects:

$$D4 = \sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right| \tag{13}$$

5) ***Chebychev distance****:* As found in Yang and Trewn (2004) it measures distance by assuming only the most significant dimension is relevant:

$$D5 = \max_j \left\{ \left| x_{sj} - x_{tj} \right| \right\} \tag{14}$$

6) ***Cosine distance:*** As defined in Tan, Steinbach, and Kumar (2005) it calculates the dissimilarity between two vectors by determining the cosine of the angle between them:

$$D6 = 1 - \frac{x_s x_t'}{\sqrt{(x_s x_s')(x_t x_t')}} \tag{15}$$

7) ***Correlation distance:*** As defined in Tan, Steinbach, and Kumar (2005) it measures the dissimilarity of the sample correlation between points as sequences of values.

$$D7 = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}}, \tag{16}$$

where $\bar{x}_s = \frac{1}{n}\sum_{j=1}^{n} x_{sj}$ and $\bar{x}_t = \frac{1}{n}\sum_{j=1}^{n} x_{tj}$

8) ***Spearman distance:*** Originally defined by Spearman (1904) to measure the dissimilarity of the sample's Spearman rank correlation between observations:

$$D8 = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}} \tag{17}$$

Where r_{sj} is the rank of x_{sj} taken over $x_{1j}, x_{2j} \ldots x_{nj}$, r_s, and r_t are the coordinate-wise rank vectors of x_s and x_t, *i.e.* $r_s = (r_{s1}, r_{s2}, .., r_{sm})$ and $\bar{r}_s = \frac{1}{m}\sum_{j=1}^{m} r_{sj} = \frac{(m+1)}{2}$, $\bar{r}_t = \frac{1}{m}\sum_{j=1}^{m} r_{tj} = \frac{(m+1)}{2}$

In our literature review (Chaudhuri and Nirupam, 1995; Cernadas *et al.*, 2005; Holalu and Arumugam, 2006; Devendran, Hemalatha, and Amitabh, 2009), most researchers compare image feature vectors as histogram-like structures. We present the following measures in terms of histograms, as they are widely used for smaller-scale image analysis in different domains.

In order to represent our feature vector as a histogram, we treated each element of n as a bin ($n = 64$). For example, we convert x_s to the histogram A, the value in each bin $[A_j]$ (for $j = 1$ to n) is equal to each x_{sj} (for $j = 1$ to n).

9) ***Hausdorff Distance:*** As defined in Munkres (1999) it measures the maximum distance from one histogram to the nearest point in the other histogram:

$$D9\,(A,B) = \max\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\} \tag{18}$$

Where sup represents the *supremum,* and inf represents the *infimum*, and *d(a,b)* represents any distance measure between two bins; we used Euclidean distance.

10) ***Jensen–Shannon divergence (JSD):*** As found in Lin (2001), it is also known as the total divergence from the average. Jensen–Shannon divergence is a symmetrized and smoothed version of the Kullback–Leibler divergence:

$$D10\,(A,B) = \sum_{j=1}^{n} A_j \log\frac{2A_j}{A_j + B_j} + B_j \log\frac{2B_j}{B_j + A_j} \tag{19}$$

11) χ^2 ***distance:*** As found in Shahrokni (2004), it measures the likeliness of one histogram being drawn from another one:

$$D11\,(A,B) = \sum_{j=1}^{n} \frac{A_j - B_j}{A_j + B_j} \tag{20}$$

12-13) ***Kullback–Leibler divergence (KLD)***: Originally defined by Kullback and Leibler (1951), it measures the difference between two histograms *A* and *B*. Often intuited as a distance metric, the KL Divergence is not a true metric since it is not symmetric, the KL Divergence from *A* to *B* is not necessarily the same as the KL divergence from *B* to *A:*

$$D12-13\,(A,B) = \sum_{j=1}^{n} A_j \log\frac{A_j}{B_j} \tag{21}$$

Since this is the only non-symmetric measure that we use for this work, we treat it as a directed measure and consider *A* to *B* and *B* to *A* as two different distances.

The last four measures (Fractional Minkowski based) are given for an *m*-by-*n* data matrix *[X]* (in our case it contains *m* = 1600 images and *n* = 64 image parameter values), which is treated as *m* (*1*-by-*n*) row vectors $[x_1, x_2, \ldots, x_m]$, the various distances between the vector x_s and x_t are defined as follows:

14-18) ***Fractional Minkowski:*** Defined as in Yang and Trewn (2004), for metrics 14 to 18, we have selected five different factional values for the Minkowski metric (0.25, 0.50, 0.80, 0.90, and 0.95). Previous research (Aggarwal, Hinneburg, and Keim, 2001; Francois, Wertz, and Verleysen, 2005) has shown that these fractional metrics out-perform the traditional Euclidean (*p* = 2) and city block (*p* = 1) distances for several artificial datasets:

$$D14-18 = \left(\sum_{j=1}^{n} \left|x_{sj} - x_{tj}\right|^{p}\right)^{1/p} \tag{22}$$

References

Aggarwal C., Hinneburg A., Keim D.A.: 2001, On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J, Vianu, V. (eds.) *Internat Conf Database*

Theory, Springer 2001, 420-434.

Banda, J.M., Angryk, R.: 2009, On the effectiveness of fuzzy clustering as a data discretization technique for large-scale classification of solar images. In: Feng, G.G. (ed.) *Proc IEEE International Conference on Fuzzy Systems*, IEEE, New York, 2019-2024.

Banda, J.M., Angryk, R.: 2010 a, An Experimental Evaluation of Popular Image Parameters for Monochromatic Solar Image Categorization. In: Guesgen, H., Murray, C. (Eds.) *The 23rd Florida Artificial Intelligence Research Society Conf*. 380-385.

Banda, J.M., Angryk, R.: 2010 b, Usage of dissimilarity measures and multidimensional scaling for large scale solar data analysis. In: Srivastava, A., Chawla, N., Yu, P., Melby, P. (eds.) *Proc 2010 Conference on Intelligent Data Understanding (CIDU) 2010,* NASA Ames Research Center 2010, 189-203.

Banda, J.M., Angryk, R.: 2010 c, Selection of Image Parameters as the First Step Towards creating a CBIR System for the Solar Dynamics Observatory. In: Zhang, J., Shen, C., Geers, G. (eds.) *Proceedings of International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE 2010, 528-534.

Banda, J.M, Angryk, R., Martens, P.C.H.: 2012, On Dimensionality Reduction for Indexing and Retrieval of Large-Scale Solar Image Data. *Solar Phys.* **283**, 113-141, doi:10.1007/s11207-012-0027-4

Beatty M., Manjunath B.S.: 1997, Dimensionality Reduction Using Multi-Dimensional Scaling for Content-Based Retrieval. In *Internat. Conf. on Image Processing 1997*. **2**. 835.

Borg, I., Groenen, P.: 2005, *Modern multidimensional scaling: theory and applications* 2nd Ed, Springer-Verlag, 145-150.

Cernadas, E., Carriön, P., Rodriguez, P., Muriel, E., Antequera, T.: 2005, Analyzing magnetic resonance images of Iberian pork loin to predict its sensorial characteristics. In *Computer Vision and Image Understanding* **98**, No. 2, 344-360.

Chaudhuri, B. B., Nirupam, S.: 1995, Texture segmentation using fractal dimension. *IEEE Trans Pattern Analysis Machine Intelligence*. **17**, No. 1, 72-77.

Datta, R., Li J., Wang, K.: 2005, Content-based image retrieval – approaches and trends of the new age. In: Zhang, H., Smith, J., Tian, Q. (eds.) *Proc of the 7th ACM SIGMM internat workshop on Multimedia information retrieval*. ACM, 253-262

Deselaers, T., Keysers, D., Ney, H.: 2008, Features for image retrieval: an experimental comparison. In *Information Retrieval.* **11**, Issue 2, 77-107.

Devendran, V., Hemalatha, T., Amitabh, W.: 2009, SVM Based Hybrid Moment Features for Natural Scene Categorization. In *International Conference on Computational Science and Engineering*. IEEE Computer Society, **1**, 356-361.

Francois D., Wertz V., Verleysen M.: 2005, Non-Euclidean metrics for similarity search in noisy datasets. In *European Symposium on Artificial Neural Networks* 2005, 27-29.

Gonzalez, R. C., Woods R.E.: 2006, *Digital Image Processing 3rd Ed*. Prentice-Hall, 100-120.

Guo, G. D, Jain, A.K., Ma, W.Y., Zhang, H.J.: 2002, Learning similarity measure for natural image retrieval with relevance feedback. *IEEE Transactions on Neural Networks*. **13**, Issue 4, 811-820

Hall, M., Frank, E., Holmes, G., Pfahringer B., Reutemann P., Witten I.H.:2009, The WEKA Data Mining Software: An Update. In: Grossman, R., Zaiane, O., Aggarwal, C., Goethals, B. (eds.) *SIGKDD Explorations*. ACM. **11**, 10-18.

Handy, B. N.; Acton, L. W.; Kankelborg, C. C.; Wolfson, C. J.; Akin, D. J.; Bruner, M. E.; Caravalho, R.; Catura, R. C.; Chevalier, R.; Duncan, D. W.; Edwards, C. G.; Feinstein, C. N.; Freeland, S. L.; Friedlaender, F. M.; Hoffmann, C. H.; Hurlburt, N. E.; Jurcevich, B. K.; Katz, N. L.; Kelly, G. A.; Lemen, J. R.; Levay, M.; Lindgren, R. W.; Mathur, D. P.; Meyer, S. B.; Morrison, S. J.; Morrison, M. D.; Nightingale, R. W.; Pope, T. P.; Rehse, R. A.; Schrijver, C. J.; Shine, R. A.; Shing, L.; Strong, K. T.; Tarbell, T. D.; Title, A. M.; Torgerson, D. D.; Golub, L.; Bookbinder, J. A.; Caldwell, D.; Cheimets, P. N.; Davis, W. N.; Deluca, E. E.; McMullen, R. A.; Warren, H. P.; Amato, D.; Fisher, R.; Maldonado, H.; Parkinson, C.: 1999, The transition region and coronal explorer. *Solar Phys.* **187**, (2), 229-260. doi:10.1023/A:1005166902804.

Holalu, S.S., Arumugam, K.: 2006, Breast Tissue Classification Using Statistical Feature Extraction Of Mammograms. *Medical Imaging and Information Science*. **23**, (3), 105-107.

Kullback, S., Leibler, R.A.: 1951, On Information and Sufficiency. *Annals of Mathematical Statistics*. **22**, (1), 79–86.

Lam, R., Ip, H., Cheung, K., Tang, L., Hanka, R.: 2000, Similarity Measures for Histological Image Retrieval. In *the 15th International Conference on Pattern Recognition 2000*. IEEE Computer Society, **2**, 2295-2298.

Lamb, R.: 2008, An information retrieval system for images from the TRACE satellite. In Master's Thesis. Montana State University, Bozeman, MT, USA.

Lin, J.: 2001, Divergence measures based on the Shannon entropy. In *IEEE Transactions on Information Theory*. **37**, (1), 145–151.

Lux, M, Savvas, A.C.: 2008, Lire: Lucene Image Retrieval – An Extensible Java CBIR Library. In *proc. of the 16th ACM International Conference on Multimedia*. ACM, 1085-1088

Munkres, J.: 1999, Topology (2nd edition). Prentice Hall, 280-281.

Naud, A.: 2001, Neural and Statistical Methods for the Visualization of Multidimensional Data. Ph.D Thesis Uniwersytet Mikolaja Kopernika w Toruniu. 84-85.

Ojala, T., Pietikainen, M., Harwood, D.: 1996, A comparative study of texture measures with classification based feature distributions. In *Pattern Recognition*. **29**, (1), 51-59.

Pentland, A.P.: 1984, Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **6**, 661-674.

Rubner, Y., Guibas, L.J, Tomasi, C.: 1997, The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA Image Understanding Workshop*. 661-668

Schroeder, M.: 1991, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W. H. Freeman. 41-45.

Shahrokni, A.: 2004, Texture Boundary Detection for Real-Time Tracking. In: Pajdla, T., Matas, J. (eds.) *European Conference on Computer Vision 2004*. Springer, **2**, 566-577.

Shepard, R.N.: 1980, Multidimensional scaling, tree-fitting, and clustering. *Science*. **210**, (4468), 390-398.

Spearman, C.: 1904, The proof and measurement of association between two things. *Am Psychol*. **15**, 72-101.

Tamura, H., Mori, S., Yamawaki, T.: 1978, Texture features corresponding to visual perception. *IEEE Trans on Systems, Man, Cybernetics*. **8**, (6), 460 - 473.

Tan, P-N, Steinbach, M., Kumar, V.: 2005, Introduction to Data Mining. Addison Wesley.

Wen-lun, C., Zhong-ke, S., Jian, F.: 2006, Traffic Image Classification Method Based on Fractal Dimension. In: Yao, Y., Shi, Z., Wang, Y., Kinsner, W. (eds.) *IEEE International Conference on Cognitive Informatics*. IEEE, **2**, 903-907.

Yang, K., Trewn, J.:2004 Multivariate Statistical Methods in Quality Management. McGraw-Hill Professional, 183-185.