

**Presenter:** Dr. David E. Thompson

**Title:** Sunspot Pattern Classification using PCA and Neural Networks (Poster)

**Authors:** T. Rajkumar (SAIC at NASA Ames), D.E. Thompson (NASA Ames), and G.L. Slater (Lockheed Martin Solar and Astrophysics Lab)

**Abstract:**

The sunspot classification scheme presented in this paper is considered as a 2-D classification problem on archived datasets, and is not a real-time system. As a first step, it mirrors the Zürich/McIntosh historical classification system and reproduces classification of sunspot patterns based on preprocessing and neural net training datasets. Ultimately, the project intends to move from more rudimentary schemes, to develop spatial-temporal-spectral classes derived by correlating spatial and temporal variations in various wavelengths to the brightness fluctuation spectrum of the sun in those wavelengths. Once the approach is generalized, then the focus will naturally move from a 2-D to an n-D classification, where 'n' includes time and frequency. Here, the 2-D perspective refers both to the actual SOHO Michelson Doppler Imager (MDI) images that are processed, but also refers to the fact that a 2-D matrix is created from each image during preprocessing. The 2-D matrix is the result of running Principal Component Analysis (PCA) over the selected dataset images, and the resulting matrices and their eigenvalues are the objects that are stored in a database, classified, and compared. These matrices are indexed according to the standard McIntosh classification scheme.

The initial dataset, provided by LMSAL, consisted of 1596 MDI images recorded during the period 05 March 2001 to 26 December 2004, representing 20 different McIntosh classes of sunspots. To create consistent training sets and matching schemes, images were first selected that were 183x183 pixels square (3 arcmin square); non-square images were rejected in the first pass. In addition, the color resolution factors in the data carried either 8-bit mapping [0-255] or full color resolution, scaled at each image. The full color resolution images were selected in preference to the 8-bit, and max-min analysis was carried out so that all the remaining images could be normalized. Solar limb images were purged; and noisy data were eliminated for this first pass. The concept 'noisy' refers to the fact that some images have a variation in background brightness due to, for example, off-center image angles. Such variation in scene brightness could confound the neural net during this initial training development, so those images were also purged from the initial test dataset. A next step would be to include a threshold determination for each image based on contrast between sunspot pattern and local background. One might eventually include a quantitative noise contrast based on variations in the full color resolution scaling for a particular image. However, it is also worth noting that images with more than one sunspot class type in the same image were kept. The result of this filtering left 308 images, now representing 15 different McIntosh classes. Images for a training dataset were selected purely randomly, but all 15 classes were included.

The use of PCA analysis helps to build the training dataset in an optimized manner, eliminating the need for a huge training dataset for the neural network. The initial classification takes about 30 milli-seconds per image, and this could be reduced considerably by adopting this algorithm in suitable hardware. The approach to sunspot classification involves creating a training data set of images that correspond to specific PCA classes. Eigenvectors (principal components) are computed for each image and the dominant eigenvalues are stored for each class. The class characteristics are updated and modified as new sunspot classification is needed for additional images, and new eigenvalue structure is incorporated to better define each class. The corresponding distribution of eigenvalues and images contribute to a *feature space*. In this feature space construction, the effect of background is also considered part of the image. For each image, variation in background brightness are monitored, and those images with too much variation (defined by the user) are purged. However, those full images kept are fully passed through the preprocessing PCA so that eigenvalues for each scene are representative of both figure and background. Once the system is initialized, the classification system projects new image eigenvalues onto the feature space. Because PCA essentially provides a distance metric through this eigenvalue structure, the system determines the closeness of a spot pattern to any given class in the feature space, and thereby selects the classification for that pattern. Additionally, because of the inherent distance metric, the system is able to recognize multiple classes of sunspots within the same image.

This initial PCA classification procedure is shown to yield an accuracy of 80-90%. As more images are included in the database, and the training set is expanded, this accuracy will improve. The initial approach based on PCA provides a preprocessing system, and provides data reduction so that certain images that might cause ambiguous classification will not be tested. The system will become more robust as it develops. Further refinement of the classification of sunspots is then based on a three-layer feed-forward neural network. To support input to the neural network, each image is divided into an equal number of 'metapixels'. The mean value of all pixel values in each metapixel is then mapped to the centroid of that metapixel. The centroid values are then provided as the inputs to the neural network. The corresponding output class values are mapped against the derived mean values within this gridwork to the neural network. The neural network training is based on an unsupervised training algorithm. The target performance for this hybrid approach to pattern classification techniques should help the user approach 100% accuracy in classification. After further development this hybrid method will be generalized for classification of solar active region observations into distinct spatial-temporal states, corresponding to various classes of flare likelihood, thus providing a sophisticated and adaptable flare prediction tool. The technique can also be applied to other areas in space weather research for pattern recognition and classification.

POC: Dr. David E. Thompson, [dethompson@mail.arc.nasa.gov](mailto:dethompson@mail.arc.nasa.gov), 1-650-604-4759