Intelligent Approach to Data Analysis with the Solar Feature Catalogues

V V Zharkova[a], J. Aboudarham[b], S Zharkov[a],
S S Ipson[a], A K Benkhalil[a] and N. Fuller[b]

[a] Department of Cybernetics, Bradford University, Bradford BD7 1DP, UK
[b] LESIA, The Paris-Meudon Observatory, F92195 Meudon Principal Cedex, FRANCE

Abstract

The searchable Solar Feature Catalogues (SFC) developed using automated pattern recognition techniques from digitized solar images are presented. The techniques were applied for detection of sunspots, active regions, filaments and line-on-sight magnetic neutral lines in the automatically standardized full disk solar images in Ca II K1, Ca II K3 and Ha taken at the Meudon Observatory and white light images and magnetograms from SOHO/MDI. The results of automated recognition were verified with the manual synoptic maps and available statistical data that revealed good detection accuracy. Based on the recognized parameters a structured database of the Solar Feature Catalogues was built on a mysql server for every feature and published with various pre-designed search pages on the Bradford University web site http://www.cyber.brad.ac.uk/egso/. The SFCs with a coverage of 10 years (1996-2005) present the most important information for the investigation of the solar feature classification and activity forecast.

1. Introduction

With a substantial increase in size of solar image data sets, the automated detection and verification of various features of interest is becoming increasingly important for, among other applications, the reliable forecast of the solar activity and space weather and data mining. However, this raises the accuracy and reliability requirements to the detection techniques applied for an automated recognition that have to be significantly improved in comparison with the existing manual ones. One of the chief objectives for European Grid of Solar Observations (EGSO) Project Work Package 5 is production of Solar Feature Catalogues by means of automated feature recognition methods. Amongst such features of interest are sunspots.

There is a growing number of archives of digitized images of the Sun, taken from ground-based and space instruments in various wavelengths. These archives are available from different locations and are to be included into a unified catalogue by the European Grid for Solar Observations (EGSO) project (Bentley, 2002). Digitized solar images from different sources have a variety of sizes, resolutions, dynamic ranges and instrumental and weather associated distortions. All are to be subjected to automated recognition processes in order to provide reliable data on the locations of features and their evolution at different times relative to solar rotation. This is aimed partly at the growing demand for solar activity forecasts by the space weather project and by many industrial organizations, which have a great need for the development of reliable and fast techniques for feature recognition on solar disks and their presentation in Solar Feature Catalogues. These catalogues are intended to contain comprehensive statistics of active events (sunspots, active regions, filaments, flares, etc.), overlapping in a given period of time and to allow the extraction of physical characteristics, which are essential for the solar activity forecast.

This requires to design advanced image recognition techniques in order to identify individual features (sunspots, active regions, filaments, magnetic neutral lines, etc.) on the images with strongly varying background caused by different terrestrial atmosphere observing conditions of solar atmosphere activity period, irregularities in shape caused by instrumental errors or any other noise in images like strips or signatures etc. For added reliability, these algorithms have to use cross-referenced criteria at multiple wavelengths in order to correctly identify the features of interest while fully utilizing all the datasets linked into the Grid.

The next following sections describe the techniques, which have been developed to date for the automated detection of sunspots, active regions, filaments and magnetic inversion lines.

2. Automated image standardization technique

There are number of difficulties that can occur with a solar image as it is demonstrated in Fig. 1: errors in FITS header information ; image shape (ellipse), error in the centre and the pole coordinates ; weather transparency (clouds) and different thickness of atmosphere ; centre-to-limb darkening ; defects in data (strips, lines, intensity). These original observation data with the header are stored in an ***observation table*** of the SFC database (see Section 4).

These distortions require to apply automated procedures for limb detection and fitting when the geometrical information provided in the image header files is not reliable and external photometric effects like limb darkening are to be removed. Robust techniques are developed to put the H_α and Ca K lines full disk images taken at the Meudon Observatory into a standardized form of a 'virtual solar image' (Ipson et al., 2003; Zharkova et al, 2003). The techniques include limb fitting, removal of geometrical distortion, centre position and size standardization and intensity normalization.

The limb fitting starts with an initial estimate of the solar centre using raw 12-bit image data and then applies a Canny edge-detection routine. Candidate edge points for the limb are selected using a histogram based method and the chosen points fitted to a quadratic function by minimizing the algebraic distance using SVD. The five parameters of the ellipse fitting the limb are extracted from the quadratic function. These parameters are used to define an affine transformation that transforms the image shape into a circle. Transformed images are generated using the nearest neighbor, bilinear or bicubic interpolation. Intensity renormalization is also required because of a limb darkening and other non-radial intensity variations. It is achieved by fitting a background function in polar co-ordinates to a set of sample points having the median intensities and by standardizing the average brightness.

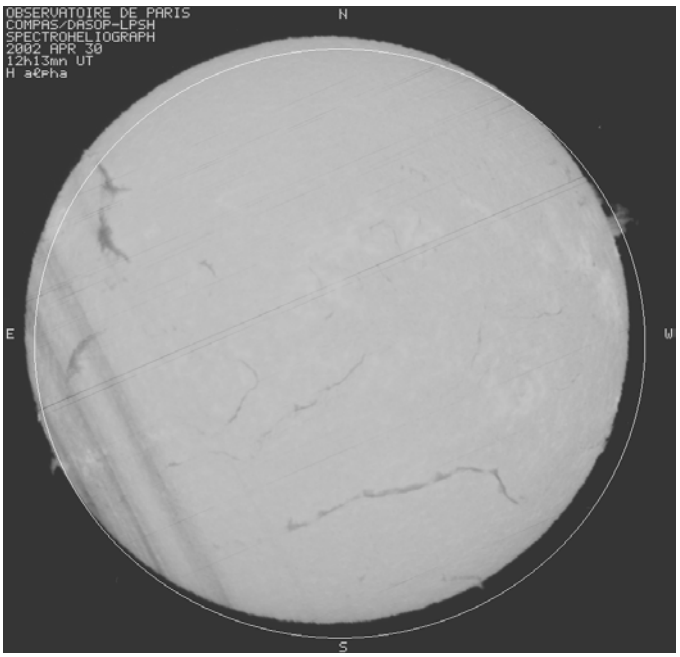

Fig. 1. A sample solar image that demonstrates the distortions in intensity (strips, limb darkening) and elliptical shape. The white line circle shows the solar disk position taken from the image header.

The parameters used for an image preprocessing, or standardization, including a code version and procedures applied are stored in in a ***pre-processing table*** of the SFC database (see Section 4).

2. Automated Feature Recognition Techniques

3.1. Sunspot detection

The solar images considered here have a "quiet Sun" background on which are superimposed bright (faculae) and dark (sunspot) features. The sunspots are features in the solar photosphere and can be observed in both white light and Ca II K1 spectral line images. Sunspots are the sites of strong magnetic field at the surface of the Sun. Visually they consist of the two parts: a dark, roughly circular central disk called the umbra, and a lighter outer area called the penumbra. Sunspots are most clearly observed in the "white light" images, but the stronger spots can also be detected in Ca II K1 images due to the absorption in that line.

Sunspot identification is required for a quantitative study of the solar cycle and this includes determining their locations, lifetimes, contrasts and other characteristics. Sunspot identification also plays an essential part in modeling of the total solar irradiance (TSI) and the variations of sunspot properties with latitude and phase in the solar cycle. Sunspots are also part of solar Active Regions, and their local behavior is used in the study of Active Region evolution and for the forecast of solar activity

Some results of sunspot detection using automated techniques developed as a part of Work Package 5 are presented in Fig. 2 (Zharkov, 2003, Zharkov, 2004a). The process involves first, if necessary, pre-processing the full disk high-resolution solar image (Fig. 2a) by correcting, if necessary, the shape of the disk to a circular one and by removing limb-darkening as described in Section 2 (Ipson et al., 2003; Zharkova et al, 2003). Then a morphological gradient operator is applied to edge enhance the image, followed by thresholding in order to detect only strong edges (Fig. 2b). After removing the limb edge, a watershed operator is applied to the binary image in order to fill the sunspot area enclosed by the edges Fig. 2c. Further median filtering is used to eliminate noise and smaller features (Fig. 2d). The regions' statistical properties are then used for the removal of false identifications such as, for example, the artefacts and lines, often present in the Meudon Observatory images. For the extraction of the area, shape, umbra/penumbra location of the detected sunspots and their basic classification region growing, local contrast and contiguity techniques are used and the results are presented in Figs. 2d and 2e.

The automated sunspot detection technique was tested on the two month's observations (April, July 2002) of Ca II K1 line images and SOHO/MDI white light images that revealed a good correlation with the manual synoptic maps. A comparison between the automated and manual detections in the Meudon Ca II K1 images was done by calculating the False Acceptance Rate (FAR) (where we detect a feature and they do not) and the False Rejection Rate (FRR) (where they detect a feature and we do not) for available observations. We introduced a Classifier Setting (CS) equal to 1 that corresponds to a total number of Sunspot Candidates detected and CS equal to 5 that represents a number of Sunspot Candidates of size greater than 8 pixel, with mean intensity lower than quiet sun intensity, principal coefficient less than 2.1 and mean absolute deviation greater than 21. As can be expected, the FFR is lowest for the classifier value of 1 that does not exceed 15.2 % from a total sunspot number detected on a day. On contrary, FAR is lowest for the value of 5 and does not exceed 8.8% from the same number of sunspots.

Hence, one can conclude that the technique applied to sunspot detection on the Meudon Ca II K1 full disk images performs very well in comparison with other earlier methods for full disk images (Chapman and Groisman, 1984, Chapman, et al., 1994)or image fragments containing sunspots (Gyôri, 1998). The technique was then applied to the SOHO/MDI white light images for whole period (1996-2003) of observations and the detected sunspot areas were compared with the average sunspot numbers from Sunspot Index Data Centre (SIDC), which confirmed a very high correlation (0.86) between these two sets (Zharkov, 2004b). These include their location, size, number of umbras and intensity range (max and min). In addition, sunspots, detected with the excellent accuracy were overlaid with the SOHO/MDI magnetograms that provided magnetic flux confined in penumbras and umbras. The sunspot parameters are extracted from the SOHO/MDI data obtained in 1996- 2003 and populated into a ***sunspot feature table*** of the SFC database discussed in Section 4.

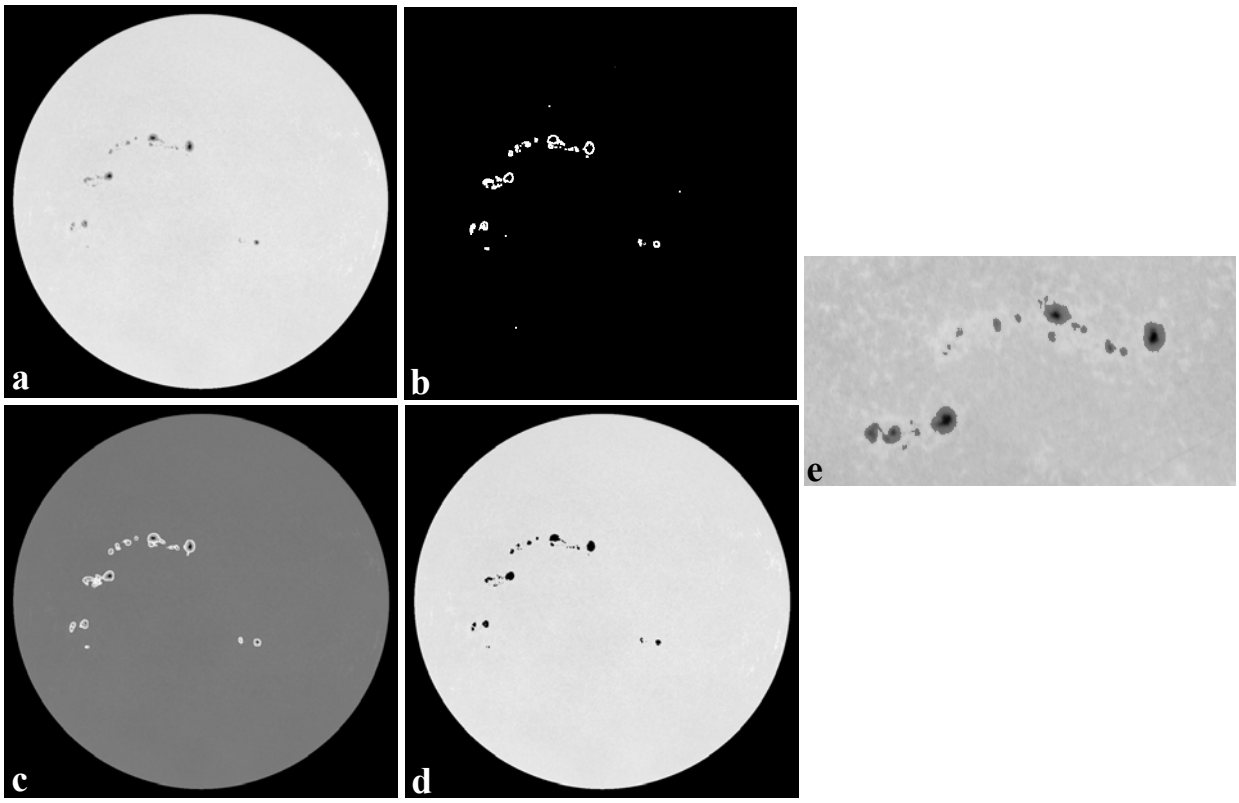

Fig. 2. Sunspot Detection performed on a Ca II K1 line full disk image obtained from the Meudon Observatory. a) The original image after cleaning, b) the final detection results superimposed on original image and c) close-up of detected umbras and penumbras in (b).

2.2. Active Region Detection

Active regions are the basic reference features for solar activity. Their reliable automated detection will enable the building of a major database of solar active features and for the first time allow analysis of solar activity on a comprehensive database of active regions taken in various wavelengths. Techniques have been developed for the automated detection of Active Regions (plages) in solar images obtained from the Meudon Observatory, using the Hα and Ca II K3 spectral lines (Benkhalil et al., 2003, Benkhalil et al., 2004), aiming to replace the existing manual detection methods (Mouradian, 1998).

The automated technique start with an initial segmentation of active regions, which is achieved using intensity thresholds determined using statistical information obtained for each quarter of a full disk solar image. Median filtering and morphological operations are applied to the resulting binary image to remove noise and to merge broken regions. Seed pixels selected in each of the initially segmented located regions are used to initiate more accurate region growing procedures. Statistically based local thresholding is applied to calculate upper and lower threshold values which control the spatial extents of the final detected regions. The technique has been tested on full-disk solar images from the Meudon Observatory for the two months of April and July 2002 and compared with their manually generated synoptic maps. Fig. 3 shows some active region detection results, Fig. 3a shows a cleaned Meudon Ca II K3 image which is the input to the procedure, Fig. 3b shows the results of remapping into the polar coordinates, Fig. 3c shows the results of the initial segmentation Fig. 3d shows the

detected regions after applying median and morphological processing and transformation back to Cartesian coordinates and finally Fig. 3e shows the final results of applying the region growing procedure superimposed on the in cleaned image. Fig. 4a shows a close-up of a detected active region Fig. 4b shows the boundary of the detected active region which is stored using a chain code (Benkhalil et al., 2003, Benkhalil et al., 2004).

A quantitative comparison was made between the results obtained using the present technique, those done manually at the Meudon Observatory (Mouradian, 1998) and those done by the National Oceanic and Atmospheric Administration Observatory (NOAA). In order to quantify the comparison, the FAR and the FRR were calculated for each day. Generally Meudon lists significantly more active regions than either us or NOAA. For most days a higher number of active regions were detected by us than by NOAA with an average FAR of 1.7 per day. The FRR of 0.2 was very low and there are only 5 days when we failed to detect a region detected by NOAA. In some cases we detect an active region while NOAA splits it into two regions. This affects the quantitative comparison. The reason for these different results is due to differences adopted for the definition of an active region. At Meudon all bright regions (plages) are detected, and these are defined as the regions in the chromosphere that are brighter than the normal "quiet" Sun background. At NOAA a detected active region is defined as a bright area on the Sun with a large concentration of magnetic field, often containing sunspots. However, not all plages contain a strong magnetic field as they might be decaying active regions with a weakening magnetic field (Driel-Gesztelyi, 2002).

The procedures developed for the automated detection of active regions have achieved a satisfactory accuracy in the detection and segmentation of active regions using full disk Hα and Ca II K3 solar images from the Meudon Observatory and full disk Fe XII 195Å solar images from SOHO. The structure of active regions at various levels of the solar atmosphere can provide a key to the understanding and reliable forecast of solar activity manifestations such as: solar flares, coronal mass ejections (CMEs), eruptive filaments etc. The parameters extracted from automatically detected active regions, such as their location, sizes, intensities (max, min and mean) are populated into the ***active region feature table*** of the SFC database discussed in Section 4.

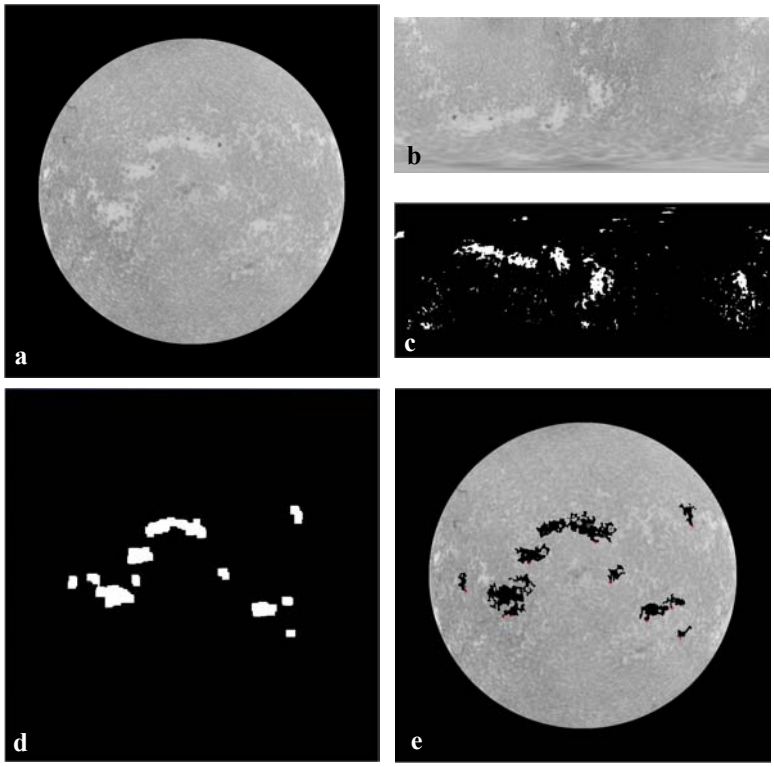

Fig. 3. The segmentation procedure stages: a) an original Ca II K3 image; b) after a transformation to Polar coordinates; c) after an initial thresholding; d) after a transformation back to the Cartesian coordinates, cleaning and morphological processing and e) a final result of the region growing.

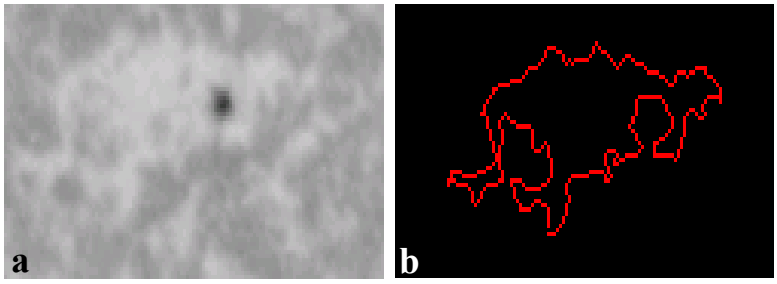

Fig. 4. a) close-up of a detected active region and b) boundary of the detected active region.

2.3. Filament Detection

Filaments are automatically detected in Hα line full disk solar images obtained at the Meudon Observatory and cleaned with our cleaning procedure (Ipson et al., 2003; Zharkova et al, 2003) using a hybrid region growing technique (Fuller, 2004). At first the seeds of filaments are to be found in order to initiate the region growing. This is achieved by enhancing a contrast of the original image by applying a linear contrast stretch to the intensity range from zero to the value which excludes the top 1% of the area of the image histogram near maximum intensity. This standardization has the effect of putting the intensities of the darker areas near 0, so that a low intensity threshold can be applied to get seeds. A high threshold value, used by the region growing process, is also calculated from the histogram.

Then one needs to calculate a size of the area of the region found when region growing is stopped using the high threshold. The first condition is that if the size is bigger than a size limit (either a fixed size or function of the seed size), then the threshold is reduced until the size condition is satisfied. A second condition is that the final threshold can not be lower than half of the original one. With these two conditions large area filaments (> limit size defined above) are retained and small ones, spread over the solar disk, are avoided. As image intensity and sharpness can vary over the solar disk, threshold values can be calculated from the neighborhood of the seed instead of from the whole image. Fig. 5 shows the filament detection results in Hα line .

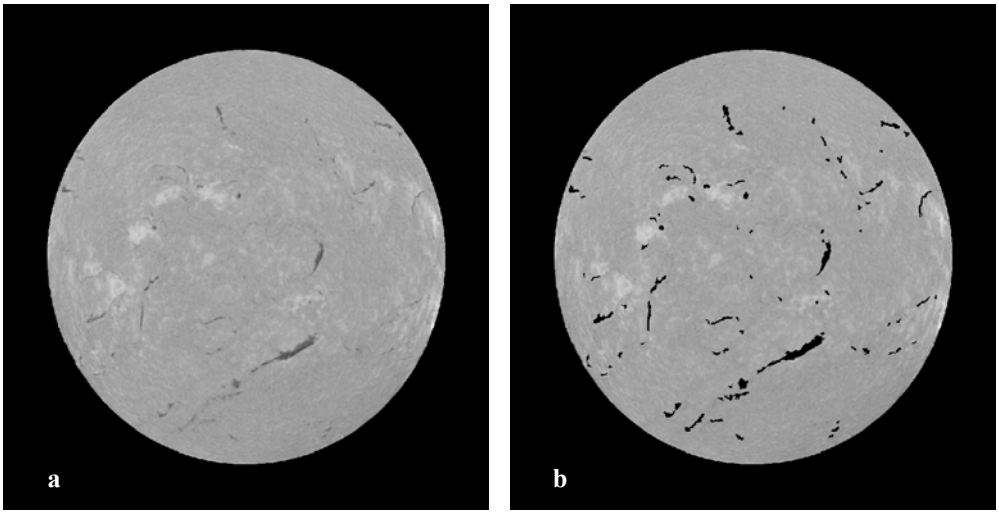

Fig. 5. Filament Regions detection in Hα line full disk images obtained at the Meudon Observatory. a) The original image after cleaning, b) the final detection results superimposed on the cleaned image.

4. Searchable database of solar features

The extracted parameters of detected features (sunspots, active regions and filaments) (Aboudarham and Zharkov, 2004) were stored as ASCII files, which are used to populate the mysql searchable databases. The databases are published on the project website http://www.cyber.brad.ac.uk/egso/SFC/SFC_form.html. They can be searched by any of the extracted parameters and downloaded in ASCII or XML formats.

The database was designed to include parameters describing the pre-processing and feature-detection code that was used for the extraction of the feature parameters as well as observational and individual feature parameters themselves.

The detection process for each can be summarized as follows:
First initial observation is pre-processed using the cleaning code (ref.) with the setup, generally, depending on the source of the observation. Features are then detected using the feature recognition methods described above. A number of parameters are then extracted for each type of the feature. Each feature is then stored as a set of pixels on a pre-processed image as either Bounding Rectangle Raster Scan for sunspots or Chain Code for filaments and active regions (faculae).

Hence, the database contains the following tables (Fig. 6):
Observation table that is related to original observation:

- Observations, which includes the observational parameters as related to the original observation;
- Observatory, which contains parameters related to the Observatory/Instrument (linked to Observations)

:
Pre-processing table that is are related to pre-processing stage:

- Pre-Processing Info - contains information about pre-processing code version, where it was run etc
- Pre-Processing Setup - which describes pre-processing code settings and input parameters.
- Pre-Processing Output - which contains the parameters which have been extracted or amended in the pre-processing stage, such as (where applicable) quiet sun intensity, image size, resolution, Solar Disk Radius,
- Ellipse fitting parameters etc.

Feature tables, or the tables related to feature recognition itself:

- Feature Recognition Code Info - provides the information describing the code used for the extraction of feature parameters.
- Feature parameters themselves (variable for different features)

Currently, there are three feature tables: Sunspots, Faculae, Filaments. Each table contains the individual feature parameters extracted from the detected feature that are described in the Feature Parameter document (Aboudarham and Zharkov, 2004).

The database structure is such that each feature is related to a processed observation (one record in Pre-Processing Output), and one Feature Recognition Code Info entry. Each processed observation is related to one original observation (Observations table), and one entry from Pre-Processing Setup and one entry from Pre-Processing info. Each original observation is associated with one entry from the Observatory table. This allows to keep the searches fully referenced to the original observations and applied standardization techniques.

5. Conclusions

In this paper we described the automated procedures used for detection of three basic solar activity features (sunspots, active regions and filaments) in the full disk solar images in Ca II K1, Hα images from Meudon Observatory and MDI from SOHO standardized by a shape and intensity. The applied techniques allowed to automatically detect the features and to extract their locations, sizes and intensities with good accuracy for the period of May 1996 – December 2003 (to be extended to 2004). These techniques allow us to build automatically synoptic maps, to overlay them and to extract the data relevant to solar activity such as features numbers, area

etc and their association with magnetic field. The extracted parameters are populated into the solar feature catalogue of sunspots, active regions and filaments that is a searchable mysql database. The searches can be performed as for single features so for their combinations.

In the future, SFC can be used for a feature classification as the solar activity flags, of every single feature and/or their combinations that will allow to find the appropriate classifiers responsible for the solar activity and to produce a short-term and long-term solar activity forecast.

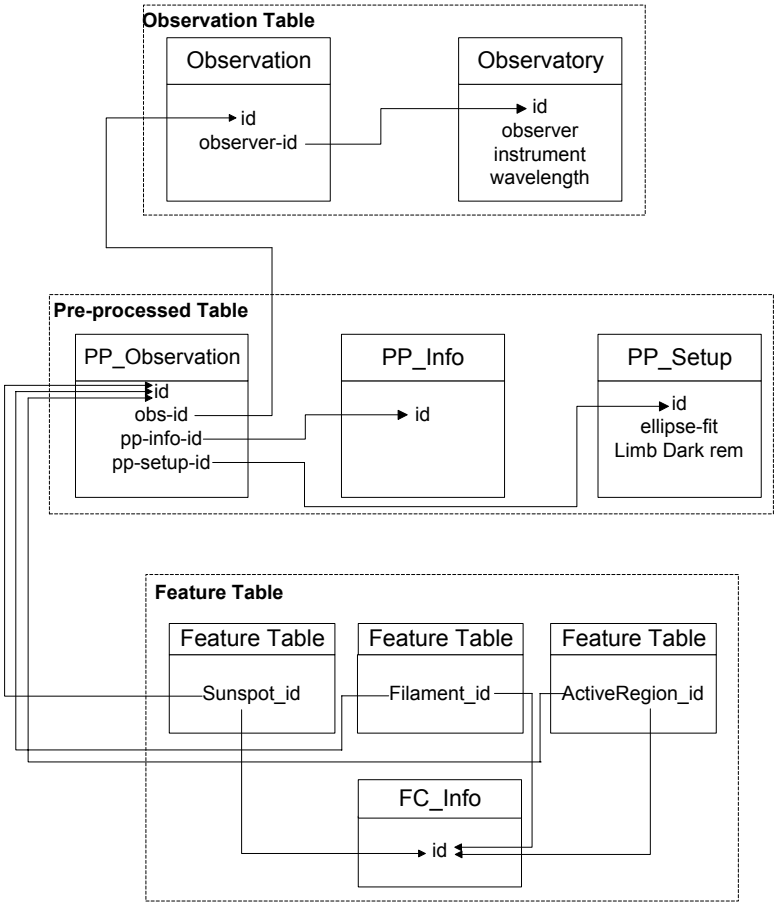

Figure 6. The SFC database structure containing the observation table, pre-processing table and feature tables for sunspots, active regions and filaments (see Section 4 for more details).

6. Acknowledgements

The authors would like to acknowledge the financial support of the European Commission funding this research as a part of the EGSO project within the IST Framework 5, project IST-2001-32409.

References

Aboudarham J and Zharkov, S.., Feature parameters document, EGSO-WP5-IR3-2.1, http://www.cyber.brad.ac.uk/egso/SFC2/FeaturesParametershtml.htm, 2004

Benkhalil, A.K., Zharkova, V., Ipson, S. and Zharkov S. Automatic Identification of Active Regions (Plages) in the Full-Disk Solar Images Using Local Thresholding and Region Growing Techniques: Proceedings of the

AISB'03 Symposium on Biologically-inspired Machine Vision, Theory and Application, University of Wales, Aberystwyth 7th - 11th April 2003, 66-73, 2003.

Benkhalil, A.K., Zharkova, V., Ipson, S. and Zharkov S. An Automated Recognition of Active Regions on the Full Disk Solar Spectroheliograms using Ha, Ca II K3 and Fe XII 195 Å Lines: Accepted to be published in the International Journal Of Computer And Their Applications, ISCA, paper ID #3303AP, 2004.

Bentley, R.D. EGSO - the next step in data analysis: Proceedings of the Second Solar Cycle and Space Weather Euro-conference, 24-29 September 2001, Vico Equense, Italy, Edited by Huguette Sawaya-Lacoste, ESA Publication SP-477, 2002.

Chapman, G.A., Groisman, G. A digital analysis of sunspot areas: Solar Phys., 91: 45, 1984.

Chapman, G.A., Cookson, A.M., Dobias, J.J. Observations of changes in the bolometric contrast of sunspots: Astrophysical Journal. 432: 403-408, 1994.

Driel-Gesztelyi, L.V., Emergence and Loss of Magnetic Flux on the Solar Surface: Proc. SOLMAG the Magnetic Coupling of the Solar Atmosphere. Euroconference and IAU Colloquium 188, Santorini,Greece, 11-15 June 2002, 113-116, 2002.

Fuller, N. Aboudarham, J. Automatic detection of Solar Filaments:Accepted to be published in the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES2004), Springer Lecture Notes in Computer Science, 2004.

Gyôri, L. Automation of area measurement of sunspots: Solar Physics, 180, 109-130, 1998.

Ipson, S.S, Zharkova, V.V, Benkhalil, A.K, Zharkov, S.I, Aboudarham J. Automated image standardization of the synoptic solar observations at the Meudon observatory: SPIE Astronomical Telescopes and Instrumentation, Astronomical Data Analysis II (AS14) Conference, 22-28 August 2002, Hawaii, USA. Innovative Telescopes and Instrumentation for Solar Astrophysics. Edited by Stephen L. Keil, Sergey V. Avakyan . Proceedings of the SPIE, 4853, 675-686, 2003.

Mouradian, Z. Synoptic Data Findings: Synoptic Solar Physics ASP Conferences Series. 140, 181-204, 1998.

Zharkov, S.I., Zharkova, V.V., Ipson, S.S., Benkhalil, A.K. An Automated Recognition of Sunspots from the Ca K1 and White Light Solar Images and Corresponding Magnetic Structures from the SOHO/MDI Magnetograms: Proceedings of the AISB'03 Symposium on Biologically-inspired Machine Vision, Theory and Application, University of Wales, Aberystwyth 7th - 11th April 2003,74-84, 2003.

Zharkov, S.I., Zharkova, V.V., Ipson, S.S., Benkhalil, A.K. Automated Recognition of Sunspots on the SOHO/MDI White Light Solar Images: Accepted to be published in the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES2004), Springer Lecture Notes in Computer Science, 2004a.

Zharkov, S.I., Zharkova, V.V. Statistical Analysis of sunspot Area and Magnetic Flux Variation in 1996-2003 Extracted from the SOHO/MDI Data: Adv. Space Res., subm., 2004b.

Zharkova, V.V., Ipson, S.S., Zharkov, S.I., Benkhalil, A.K., Aboudarham, J. Bentley, R.D. A full disk image standardization of the synoptic solar observations at the Meudon observatory: Solar Physics Journal, 214/1, 89-105, 2003.