Automatic Detection of Active Region on EUV Solar Images Using Fuzzy Clustering

M. Carmen Aranda and Carlos Caballero

Department of Languages and Computer Science,
Engineerings School,
University of Malaga,
C/ Doctor Ortiz Ramos s/n, 29071 Málaga, España (Spain)
mcarmen@lcc.uma.es,carlos.caballero.gonzalez@uma.es

Abstract. The technique presented in this paper is based on fuzzy clustering in order to achieve robust automatic detection of active regions in solar images. The first part of the detection process is based on seed selection and region growing. After that, the regions obtained are grouped into real active regions using a fuzzy clustering algorithm. The procedure developed has been tested on 400 full-disk solar images (corresponding to 4 days) taken from the satellite SOHO. The results are compared with those manually generated for the same days and a very good correspondence is found, showing the robustness of the method described.

Key words: fuzzy clustering, cluster validity measure, active region detection, image processing, region growing

1 Introduction

The automatic processing of information in Solar Physics is becoming increasingly important due to substantial increase in the size of solar image data archives and also to avoid the subjectivity that carries the manual treatment of this information. The automated detection of solar phenomena such as sunspots, flares, solar filaments, active regions, etc, is important for, among other applications, data mining and the reliable forecast of the solar activity and space weather. Significant efforts have been done to create fully automated Solar Catalogues[1].

In this paper we focus on the automatic detection of active regions on solar Extreme Ultraviolet (EUV) images obtained from the satellite SOHO. Active regions are solar regions with intense magnetic activity which can be detected as bright regions in the bands of $H\alpha$ or EUV. It could be useful to study its evolution and behaviour in the forecast of solar flare activity. Active regions have been manually detected and numbered for dozen of years by the NOAA (National Oceanic and Atmospheric Administration) organization. Some automated detection methods have been developed in order to avoid the inherent subjectivity of manual detections[2, 3].

As region growing has proved to be a reliable means to investigate solar features as filaments[3] or active regions, we have based our method on it, improving the way seeds and thresholds are chosen. After that, fuzzy clustering is applied to group the candidate regions produced into complete active regions. Fuzzy clustering allows the introduction of fuzziness for the belongingness of each bright region to a concrete active region. Fuzzy clustering has been widely used in image processing for image segmentation or boundary detection, also for solar images[4, 5].

The rest of the paper is organized as follows. Section 2 presents the preprocessing stage prior to the region selection. Section 3 introduces the seed selection and region growing procedures which produce a bright regions automatically selected candidates to belong to real active regions. This is followed by the description of the fuzzy clustering procedure in Section 4. A validity measure is also defined to choose the optimal number of clusters in the image. Section 5 shows some experimental results. Finally, Section 6 summarizes the main conclusions of the paper.

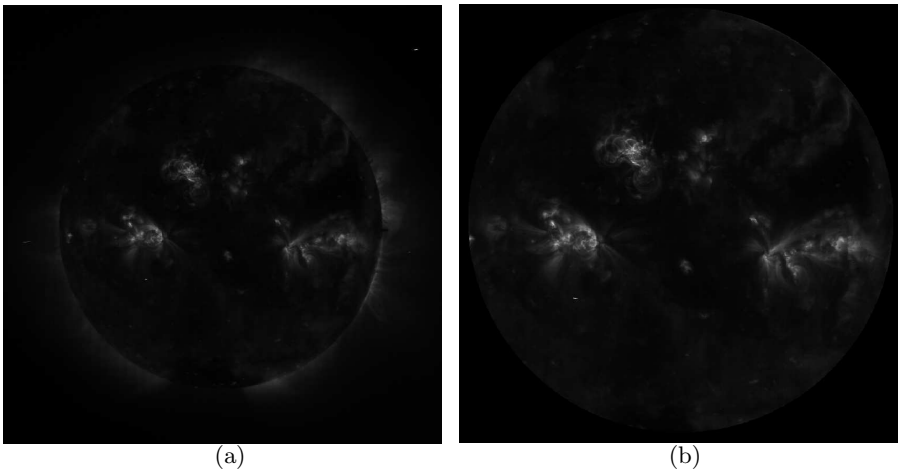
2 Image Preprocessing

Prior to any feature recognition, the solar images have to be pre-processed in order to correct them from geometrical or photometric distorsions. The images used for the process are Extreme Ultraviolet (EUV) images of the Sun acquired from the satellite SOHO (Solar and Heliospheric Observatory). They are downloaded in FITS (Flexible Image Transport System) file format. FITS is the most commonly used digital file format in astronomy. It is designed specifically for scientific data and hence includes many descriptions of photometric and spatial calibration information and image origin metadada. The calibrations applied to the images were:

- **Dark current subtraction:** a uniform (identical for all the pixels) zero flux response is subtracted from the raw image.
- **Degridding:** the aluminum filter located close to the focal plane of the instrument casts a shadow on the CCD detector that creates a modulation pattern, or *grid*, in the images. The degridding factors are calculated and stored, and the image is multiplied by the degridding factor for a fairly reasonable correction to the data.
- **Filter normalization:** account is taken for the variable transmittivity of the clear and aluminum filters (Al+1 or Al+2).
- **Exposure time normalization:** the flux is normalized to the exposure time. Binned images are treated properly.
- **Response correction:** due to exposure to EUV flux, the pixel to pixel sensitivity (flat-field) of the CCD detector is highly variable. The flat-fields needed to correct the images are computed regularly from images of visible light calibration lamps.

After the calibration on the image has been made, the background, the halo and the contour of the image are completely erased. The contour is not important

because the information received from the satellites about the contour is not completely true. This information is not true for problems capturing images of the satellites. The results can be seen in Figure 1, where the image 1 (a) is the



(a) (b)

Fig. 1: (a) original image (b) image without background, halo and contour.

original image of the Sun taken on January 15, 2005 and the image 1 (b) is the image without background, halo and contour.

3 Region Detection

Once the image is fully cleaned and pre-processed, we can investigate the active regions using first a region growing method based on the image grey level properties. The principle is to group pixels into large regions if these pixels fall into a predefined intensity range. The procedure is started from a pixel or small region called a seed. This method is more effective than applying a basic automatic threshold as it associates a grey level condition with a connectivity condition. The efficiency of the method will thus depend on the seed selection procedure and on the intensity range definition. The region growing process usually produces a big amount of regions that need to be grouped into real active regions as it will be shown in Section 4.

3.1 Seeds Selection

The seed selection is a major step in the procedure. Firstly, the method calculates the Otsu's optimal value only for the pixels of the Sun area. Let's assume that an image has 2 types of pixels: objects and background. The threshold is obtained

minimizing the weighted within-class variance. This turns out to be the same as maximizing the between-class variance.

$$Otsu's\ threshold = \max_{1 \leq t \leq L} \left\{ \sigma_W^2(t) \right\} \tag{1}$$

where

- An image contains N pixels whose gray levels are between 1 and L.
- $p_i = \frac{f_i}{N}$, f_i is the frequency of occurrence of the value i.
- Pixels are divided into two classes: $C1$, with gray levels $[1, \ldots, t]$ and $C2$, with gray levels $[t+1, \ldots, L]$.

The weighted within-class variance is:

$$\sigma_W^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \tag{2}$$

Where the class probabilities are estimated as:

$$q_1(t) = \sum_{i=1}^{t} P(i) \qquad\qquad q_2(t) = \sum_{i=t+1}^{L} P(i) \tag{3}$$

And the class means are given by:

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)} \qquad\qquad \mu_2(t) = \sum_{i=t+1}^{L} \frac{iP(i)}{q_2(t)} \tag{4}$$

Finally, the individual class variances are:

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)} \qquad\qquad \sigma_2^2(t) = \sum_{i=t+1}^{L} [i - \mu_2(t)]^2 \frac{P(i)}{q_2(t)} \tag{5}$$

The optimal value of Otsu will be used to select the seeds. Every pixel which intensity value is less than the value obtained from the Otsu's method will be seed.

$$\forall x, y \in D, seed_{x,y} = Otsu's\ threshold > I(x, y) \tag{6}$$

where

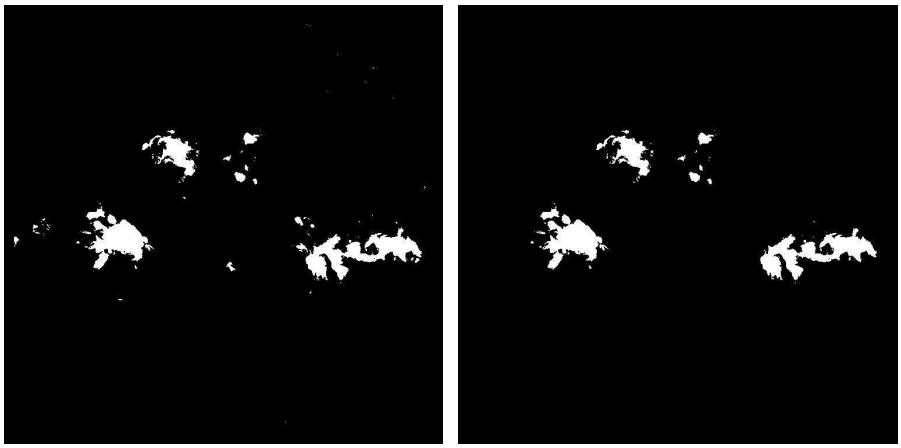
- $D \in [1 \ldots width\_Image, 1 \ldots length\_Image]$.
- $I(x, y)$ is the pixel's value on its coordinates x,y.

3.2 Region Growing

The next step consists in growing the region. The region growing code that has been used in this system was developed by Gonzalez[6]. This method uses three inputs:

- An image.
- A set of seeds (calculated as in the previous section).
- A threshold ($threshold_{limit}$) which set the limit to growth. This limit is in the range: $[seed - threshold_{limit}, seed + threshold_{limit}]$.

$Threshold_{limit}$ is calculated as the average of the values of all pixels in the image. The result obtained after applying these techniques can be seen in Figure 2 (a).



(a) (b)

Fig. 2: (a) image after region growing (b) image after selection of candidate regions

3.3 Selection of Candidate Regions

As it can be seen, after making the region growing technique there is too much noise and portions of regions that do not have the importance that they really deserve such as you can see in Figure 2. Thus, it is advisable to make a selection phase of candidate regions. This phase consists of two steps:

1. Removing all regions that do not exceed a minimum value of area. This will eliminate false positives in the selection of seeds.
2. Once large regions have been selected, the next step is to look for regions close to these, regardless of the size of these surrounding regions. Finally we get more compact regions suitable for use in the next step. The result of the segmentation process is positive, as can be seen by comparing 1 (a) and 2 (b).

4 Fuzzy Clustering Algorithm to Identify Active Regions

Once the image has been segmented into independent pieces, a grouping process should be performed for the candidate regions are clustered in real active regions. The algorithm used here is the Gustafson-Kessel fuzzy clustering algorithm[7]. Babuska[8] provided a slight variation of the Gustafson-Kessel algorithm which improved the variance estimation.

This technique has been selected for this problem because it is not known a priori the form or structure of the clusters. Another important point is that the distance measure is the distance of *Mahalanobis*[9]. Gustafson and Kessel extended the standard fuzzy c-means algorithm.

The Fuzzy c-means clustering algorithm[10] is based on the minimization of an objective function called *c-means* functional. It is defined as:

$$J(X;U,V) = \sum_{i=1}^{c}\sum_{k=1}^{M}(\mu_{ik})^m||x_k - v_i||_A^2 \tag{7}$$

where

- $X = \{x_1, x_2, \ldots, x_M\}$ are the data which must be classified.
- $U = [\mu_{ik}] \in M_{fc}$, is a fuzzy matrix of X.
- $V = [c_1, c_2, \ldots, c_c]$, is the vector of centroids.

Gustafson and Kessel extended the standard fuzzy c-means algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set. Each cluster has its own norm matrix A_i. The matrices A_i are an optimization variable in the following objective function:

$$J(Z,U,V,A) = \sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{i=k}^m(z_k - v_i)A_i(z_k - v_i)^T \tag{8}$$

But the objective function cannot be minimized directly because J depends linearly on A_i. Therefore A_i is constrained by: $det|A_i| = \rho_i, \rho > 0$. Allowing the matrix A_i to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume is constant. Using the Lagrange multiplier method, the following expression for A_i is obtained:

$$A_i = [\rho_i det(\mathbf{F}_i)]^{1/n}\mathbf{F}_i^{-1} \tag{9}$$

where $\mathbf{F}_i$ is the fuzzy covariance matrix of the ith cluster defined by:

$$\mathbf{F}_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m(\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^m} \tag{10}$$

Clustering algorithm is fuzzy does not mean that a priori information of the number of clusters to be made should not provide. To solve this problem, N iterations with N possible numbers of clusters are made in our system. The value of N is in the range of $[2, M]$. If the number of candidate regions is less than M, the number of possible clusters that exist and therefore iterations will be the number of candidate regions. If the number of candidate regions is greater than M, the number of iterations will be M. This is developed so for several reasons.

- The empirical value of M will be more or less 10. This is because is difficult to find more than ten active regions simultaneously in the Sun.
- The computational cost required to perform clustering is very high and must be optimized.

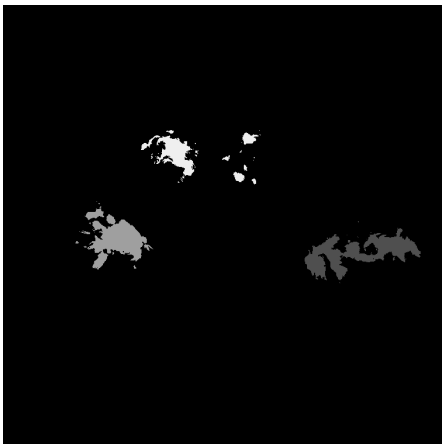


Fig. 3: Three clusters are determined in this image

4.1 Cluster Validation

Once calculated the candidate active regions for different number of clusters, it will be possible to determine the correct number of clusters. To do this we use a modified clustering validation index with different densities developed by Chou, Su and Lai[11]. The index used is defined in the Equation (11) where

- A_i is the set whose elements are the data points assigned to the ith cluster.
- $|A_i|$ is the number of elements in A_i.
- $\underline{x}_j$ and $\underline{x}_k$ are the regions' centroids.
- $\underline{v}_i$ and $\underline{v}_j$ are the clusters' centroids ith and jth.
- d is a distance function.

$$CS(c) = \frac{\displaystyle\sum_{i=1}^{c} \left\{ \frac{1}{|A_i|} \sum_{\underline{x}_j \in A_i} \max_{\underline{x}_k \in A_i} \left\{ d(\underline{x}_j, \underline{x}_k) \right\} \right\}}{\displaystyle\sum_{i=1}^{c} \left\{ \min_{j \in c, j \neq i} \left\{ d(\underline{v}_i, \underline{v}_j \right\} \right\}} \tag{11}$$

The correct number of clusters is one that minimizes the value of the index. The main difference between this method and our method is the distance function $d(\underline{x}_i, \underline{x}_k)$. In both cases the distance used is the Manhattan distance. In the first case the Manhattan distance is calculated between regions' centroids and in the second case it is calculated between the closest points of two regions. The Gustafson and Kessel algorithm determines which elements belong to each cluster and the index determines the number of clusters. Figure 3 shows the tree clusters finally obtained for the image that appears in Figure 1 (a).

5 Experimental Results and Discussion

The procedure for automatic detection of active regions has been applied to the period between January 15, 2005 and January 18, 2005, on more than 400 observations.

A summary of the results can be seen in Table 1 showing the percentages of candidate region detection and the percentage of errors. The errors are classified in ± 1 error, ± 2 errors and > 2 errors. For example, if an image has three brilliant regions and the method detects two or four candidates region it will be an error classified in the group ± 1.

			Errors		
Day	Observations	%Candidate region detected	% ± 1	% ± 2	% > 2
January 15, 2005	95	93.685	6.315	0	0
January 16, 2005	109	89.910	7.339	1.834	0.917
January 17, 2005	94	57.448	29.787	9.574	3.191
January 18, 2005	110	37.275	17.272	18.181	27.272

Table 1: Experimental results

In this study, it is taken into account that a brilliant region of the Sun is only manually numbered by the NOAA as an active region when it has high activity during its life (normally several days). So, it is not possible to determine if a brilliant region will be numbered or not as an active region considering only the information of one image. For us, all the brilliant regions will be candidate active region because they present high activity in that particular moment. The experimental results show the correspondence of the clustering output with the candidate active region that there are in each image.

A further processing should be done to analyze which candidate active regions are real active regions studying its behaviour during a long period of time.

To study the fuzzy clustering algorithm is regarded as being correctly classified all images correctly detected the active and candidates. Next, the experimental results will be discussed in detail.

5.1 January 15, 2005 and January 16, 2005

The results obtained on January 15, 2005 using a total of 95 images are:

- 93.685% of automatically detected candidate regions match the manually detected ones.
- 6.315% of automatically detected candidate regions don't correspond to manual ones.

On this day there are 2 active regions but the reality is that after the process of image segmentation we can clearly see three candidate active regions. One of

these images has been developed throughout this paper and the final result can be seen in Figure 3. The result of fuzzy clustering algorithm is quite positive of success getting a value of 93.685%.

The results obtained on January 16, 2005 using a total of 109 images are:

- 89.910% of automatically detected candidate regions match the manually detected ones.
- 7.339% of automatically detected candidate regions has ± 1 error.
- 1.834% of automatically detected candidate regions has ± 2 errors.
- 0.917% of automatically detected candidate regions has > 2 errors.

On this day the 2 regions of the previous day remain. The results remain very positive being the value of automatic detection 89.910%.

5.2 January 17, 2005

The results for the January 17, 2005 using 94 images are:

- 57.448% of automatically detected candidate regions match the manually detected ones.
- 29.787% of automatically detected candidate regions has ± 1 error.
- 9.574% of automatically detected candidate regions has ± 2 errors.
- 3.191% of automatically detected candidate regions has > 2 errors.

On this day 2 new active regions appear. So, there exist a total of 4 active regions. The percentage of failing has increased which is a more modest result than those obtained previously. Note that almost 30% of failures belong to the group ± 1.

5.3 January 18, 2005

The results for the January 18, 2005 using 110 images are:

- 37.275% of automatically detected candidate regions match the manually detected ones.
- 17.272% of automatically detected candidate regions has ± 1 error.
- 18.181% of automatically detected candidate regions has ± 2 errors.
- 27.272% of automatically detected candidate regions has > 2 errors.

On this day 6 regions are active but is a day especially complicated. Visually it is imposible to determine whether there are really six active regions because they are very close, even some of them are overlapping. At this point, we can state that the clustering algorithm works well even in these cases. To solve the problem it would be neccesary to combine information from several frequency bands.

6 Conclusions

In this paper some experimental results for the detection of active regions on the Sun have been presented using the Gustafson-Kessel's fuzzy clustering algorithm and the Mu-Chun's index validation. Satisfactory results have been obtained as have been shown in Table 1.

One of the major constraints that have been found in the system is inherent to the algorithm of Gustafson-Kessel, where the density of points per cluster is predetermined.This fact provoke a serious problem in the system because a region of large area is considered equally influential than a small region. A proposal for improving the system could be to modify the parameter input to the Gustafson-Kessel algorithm, as the division of large regions into small regions.

So, the automated techniques developed allow to detect bright regions on the Sun and to group them into real active regions. Further work will be carried out to produce automatic AR detection in other spectral bands or in the magnetogram to combine all the information to obtain more realistic detections.

This work was funded by the project TIC07-02861 of the Junta de Andalucía.

References

1. Zharkova, V., Abourdarham, J., Zharkov, S., Ipson, S., Benkhalil, A.: Searchable solar feature catalogues. In Advances in Space Research, 36, pp. 1604–1612 (2005)
2. Benkhalil, A., Zharkova, V., Ipson, S., Zharkov, S.: Active region detection and verification with the solar feature catalogue. In: Solar Physics, 235, pp. 87 (2006)
3. Qahwaji, R., Colak, T.: Automatic detection and verification of solar features. In International Journal of Imaging System and technology, vol. 15, Issue 4, pp. 199-210
4. Banda, J. M., Angryk, R.A.: On the effectiveness of Fuzzy Clustering as a data discretization technique for large-scale classification of solar images. In: IEEE International Conf. on Fuzzy Systems, pp. 2019–2024, (2009)
5. Barra, V., Delouille, V., Hechedez, J.: Segmentation on extreme ultraviolet solar images using a multispectral data fusion process. In: IEEE International Conf. on Fuzzy Systems, pp 1–6, (2007).
6. Gonzalez, R., Woods, R.: Digital Image Processing Using MATLAB. Pearson Prentice-Hall, New Jersey (2008)
7. Gustafson, E., Kessel, W.: Fuzzy Clustering with a Fuzzy Covariance Matrix. In: IEEE CDC, pp. 761–766. IEEE Press, San Diego (1979)
8. Babuska, R., Van der Venn, P.J., Kaymak, U.: Improved variance estimation for Gustafson-Kessel clustering. In: Proceedings of 2002 IEEE International Conference on Fuzzy Systems, pp. 1081–1085, Honolulu, Haway (2002).
9. Mahalanobis, P.: On the generalised distance in statistics. In: Proceedings of the National Institute of Science of India 12. pp. 49–55 (1936)
10. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algoritms. In: Plenum Press, New York (1981)
11. Chou, C. H., Su, M. C., Lai, E.: A new cluster validity measure for clusters with different densities. In: IASTED International Conf. on Intelligent Systems and Control, pp. 276–281, Austria (2003)