Images sub-segmentation with the PFCM clustering algorithm

B. Ojeda-Magaña *, J. Quintanilla-Domínguez †, R. Ruelas * and D. Andina †.
*Departamento de Ingeniería de Proyectos-CUCEI
Universidad de Guadalajara.
José Guadalupe Zuno No. 48. C.P. 45101 Zapopan, Jalisco, México.
†Departamento SSR E.T.S.I Ingenieros de Telecomunicación
Universidad Politécnica de Madrid.

***Abstract*— In this work we propose a method for sub-segmentation of images using the PFCM clustering algorithm. The sub-segmentation consists of finding, within the clusters found using the segmentation process, those data less representative, or atypical data, belonging to the clusters. These data represent, in many cases, the zones of interest during image analysis. Two different examples are used in order to show the results, and the advantages of identifying those elements of data forced to belong to a cluster, of which they are the less representative and, therefore may contain information of great interest in particular applications.**

Keywords: Image processing, Fault detection, diagnostics and prognostics.

I. Introduction

Image segmentation is an important task in the fields of image processing and computer vision. The main objective of image segmentation is to find objects or regions with the same characteristics. Several segmentation methods exist [1], [2], such as edge detection, region growth, histogram thresholding, clustering, neural networks, etc. The objective of the clustering process is to find pixels groups with a similar gray intensity, or more or less homogeneous groups. The similarity is evaluated according to a distance measure between the pixel and the prototypes of the objects or regions, and each pixel is assigned to the nearest or most similar prototype. However, this process must distribute all data to the different groups, even if some pixels are not very representative of the group as a whole.

For this reason, in many cases it is not sufficient to identify the groups or objects of an image, but to distinguish, through the variations in intensity, the imperfections present in an image. For the two examples considered in this work these imperfections could be the result of a light reflection over the image but, and this second case is a very interesting one, this could be real imperfect characteristics that can be used aid in image analysis, e.g in diagnosis.

Atypical data are generally difficult to detect because they are present in only a few pixels, and the traditional process to increase the number of clusters can force us to use a great quantity of cluster in order to detect the atypical data. A different approach, as proposed in this work, is to find the normal groups in the image first and, inside them, to find sub-groups containing atypical pixels of the different groups, looking for imperfect characteristics in the image.

In a variety of applications, based on images to detect the most zones of highest risk, such as the cancer risk analysis, the less representative data, or that tend to fall outside of a normal pattern, are precisely the most interesting because they represent a variation with respect to healthy tissue. We therefore propose image sub-segmentation as an alternative to identify this kind of data after an image is segmented in an acceptable way.

For the image sub-segmentation we propose using the *Possibilistic Fuzzy c-Means* (PFCM) [3] clustering algorithm, which finds pixels with a similar intensity level of gray and groups them together into a determined number of classes or objects. For the two cases presented in this work, the images are represented in gray levels and, particularly, we use only one characteristic corresponding to the intensity level of each pixel.

The purpose of this work is to find sub-clusters utilizing of the advantages offered by the PFCM clustering algorithm, which has the qualities of both the *Fuzzy c-Means* (FCM) [4] and the *Possibilistic c-Means* (PCM) [5]. The FCM help us to create groups of pixels, whereas the PCM helps us to identify sub-groups, or atypical data, inside each group. We use the typicality values and each group is divided in two sub-groups, the sub-group of typical data or data with typicality values greater than a specified threshold. The other sub-group contains the atypical data or data with typicality values below the established threshold.

The next section presents the PFCM algorithm. The Section III contains the results of segmentation and sub-segmentation of two kinds of images. The Section IV contains the main conclusions of this work.

978-1-4244-3760-3/09/$25.00 ©

II. Clustering algorithms

In this work we take advantage of the qualities of the fuzzy and possibilitic clustering algorithms in order to find c groups in a set of unlabeled data $Z = \{z_1, z_2, \ldots, z_k, \ldots, z_n\}$ in a M-dimensional space, that is, the nearest data z_k to a prototype, or group center, v_i, belong to this group. The membership of each data z_k to the different groups depends on the kind of partition of the M-dimensional space where data are defined. This way, a partition can be either: strict, fuzzy, or possibilistic.

The strict partition of the space for a data set $Z(k) = \{z_k | k = 1, 2, ..., N\}$, of finite dimension and c center of groups, where $2 \leq c < N$, is defined by (1), (2) defines the fuzzy partition, whereas (3) defines the possibilistic partition.

$$M_{hc} = \left\{ \mathbf{U} \in \Re^{c \times N} | \mu_{ik} \in \{0,1\}, \forall i \text{ and } k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\}; \quad (1)$$

$$M_{fcm} = \left\{ \mathbf{U} \in \Re^{c \times N} | \mu_{ik} \in [0,1], \forall i \text{ and } k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\} \quad (2)$$

$$M_{pcm} = \left\{ \mathbf{U} \in \Re^{c \times N} | \mu_{ik} \in [0,1], \forall i \text{ and } k; \forall k, \exists i, \mu_{ik} > 0; \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\}. \quad (3)$$

A. Fuzzy c-Means algorithm

The Fuzzy c-Means (FCM) [4], [6] is an algorithm that calculates a membership degree for each point (z_k) for each of n different groups $A_i, i = 1, \ldots, c$. The sum of the membership degrees of a point must be equal to one. However, a problem arises when there are several equidistant points from the center of the groups, because the FCM is not able to detect noise or nearest and furthest data from the prototypes. Pal *et al* [7] show an example with two points of data points in the boundary of two groups, one point near to the prototypes and the other one far away from them. This must be handled with care, as both points are not equally representative of the groups, even if they have the same membership degrees. One way to overcome this inconvenience is to use a possibilistic algorithm.

B. Possibilistic c-Means algorithm

The Possibilistic c-Means clustering algorithm (PCM) [5] is based on typicality values and relaxes the constraint of the FCM concerning the sum of membership degrees of a point to the n groups, which must be equal to one. Thus, the PCM identifies the similarity of data with a given number or prototypes using a typicality values that takes values in [0,1]. The nearest data to the prototypes are considered typical data, further data are atypical and data with zero, or almost zero, typicality values are considered noise.

To avoid an initial problem with this algorithm, as sometimes the prototypes of different groups coincided [8], even if the natural structure of data has well delimited different groups, Tim *et al* [9]–[11] have modified the objective function to include a constraint that there is a repulsion between the groups, thus avoiding identical groups when they must be different.

C. PFCM clustering algorithm

Pal *et al*. [12] have proposed to use the membership degrees as well as the typicality values, looking for a better clustering algorithm. They called it *Fuzzy Possibilistic c-Means* (FPCM). However, the sum equal to one of the typicality values for each point was the origin of a problem, particularly when the algorithm uses a lot of data.

In order to avoid this problem, Pal *et al* [3] proposed to relax this constraint and they developed the PFCM clustering algorithm, where the function to be optimized is given by (4)

$$\mathbf{J}_{pfcm}(\mathbf{Z}; \mathbf{U}, \mathbf{T}, \mathbf{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (a\mu_{ik}^{m} + bt_{ik}^{\eta}) \times \|z_k - v_i\|_A^2 + \sum_{i=1}^{c} \delta_i \sum_{k=1}^{N} (1 - t_{ik})^{\eta}, \quad (4)$$

and subject to the constraints $\sum_{i=1}^{c} \mu_{ik} = 1 \forall k$; $0 \leq \mu_{ik}, t_{ik} \leq 1$ and the constants $a > 0$, $b > 0$, $m > 1$ and $\eta > 1$. The parameters a and b define a relative importance between the membership degrees and the typicality values. The parameter μ_{ik} in (4) has the same meaning as in the FCM. The same happens for the t_{ik} values with respect to the PCM algorithm.

Theorem PFCM [3]: If $D_{ikA} = \|z_k - v_i\|_A > 0$, for every i, $k, m, \eta > 1$, and Z contains at least c different patterns, then $(U, T, V) \in M_{fcm} \times M_{pcm} \times \Re^{c \times N}$ and $\mathbf{J}_{pfcm}$ can be minimized if and only if

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{D_{ikA}}{D_{jkA}} \right)^{2/(m-1)} \right)^{-1} \quad (5)$$
$$1 \leq i \leq c; \quad 1 \leq k \leq N$$

$$t_{ik} = \frac{1}{1 + \left(\frac{b}{\gamma_i} D_{ikA}^2 \right)^{1/(\eta-1)}} \quad (6)$$
$$1 \leq i \leq c; \quad 1 \leq k \leq N$$

$$v_i = \sum_{k=1}^{N} (a\mu_{ik}^{m} + bt_{ik}^{\eta}) z_k \Big/ \sum_{k=1}^{N} (a\mu_{ik}^{m} + bt_{ik}^{\eta}), \quad (7)$$

$$1 \leq i \leq c.$$

The membership degrees are calculated with equation (5), the typicality values with (6) and for the prototypes the equation (7) is used.

The PFCM clustering algorithm uses four parameters, (a, b, m, η), where m defines the fuzziness level of the partition, η defines the possibilistic level of the partition, whereas a and b weigh the relative importance between the fuzzy and the possibilistic approaches. Even if the parameters m and η can take any value in a wide interval, it is classical to use 2 for both parameters. The relative value between a and b offer the option that the results (the prototypes for example) of the PFCM depend more on the fuzzy values or on the typicality values. If a is greater than b, the fuzzy values have a more important weight on the calculus. On the other hand, if b is greater than a, then the typicality values have a more important influence on the results. In order to reduce the noise effects, this last relation must be applied, that is, the value of b must be greater than a [3].

Using the fuzzy membership and typicality values it is possible to identify the groups and the most representative data. We therefore use them but we are interested in the atypical data instead of the typical data. With these atypical data we build the sub-groups.

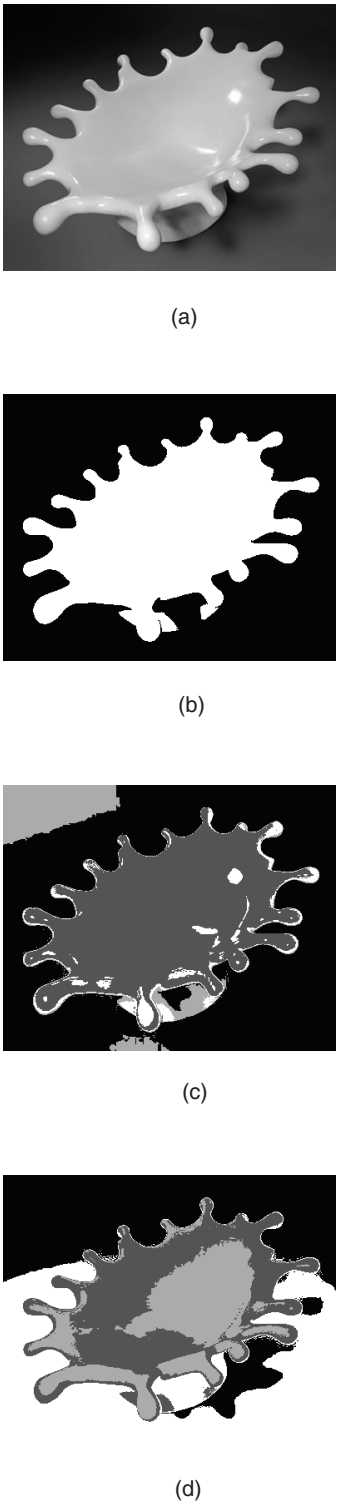
III. Segmentation with the PFCM algorithm

In this section we apply the PFCM algorithm to the image segmentation. Specifically, we use the membership μ_{ik} and the typicality values t_{ik}, as a way to find more information directly related with data that are dissimilar to the groups found in the image. We use two kinds of images, a drop of milk, see Fig. 1(a), and a mammogram, see Fig. 2(a) and Fig. 2(c). All the images are given in gray levels.

For the image segmentation we use only the intensity level of pixels as a characteristic to train the algorithm. The selected values for the parameters of the algorithm, for all the images, are: $a = 1$, $b = 2$, $m = 2$, $\eta = 2$. In the drop of milk image two classes can easily be identified, one representing the drop of milk and the other the background. Applying the PFCM algorithm for the identification of two classes, they remain represented by the membership matrix U of membership degrees and the typicality matrix T of typicality values. The Table 1 contains the values of the prototype centers for both classes.

TABLE I
Calculated prototypes with the PFCM algorithm for the image of Fig. 1(a).

PFCM (a=1,b=2,m=2,η=2)	
v1	v2
187.6164	49.0242



(a) (b) (c) (d)

Fig. 1. (a) Drop of milk image. Segmentation with the PFCM: (b) in two groups with the membership matrix U, (c) in two groups and two sub-groups with the matrix T, and (d) in four groups with the membership matrix U.

Each pixel z_k of an image produces one vector μ_{ik} in the membership matrix U, and another t_{ik} in the typicality matrix T. The membership matrix U has the constraint that the sum of each vector μ_{ik} must be equal to one as previously specified.

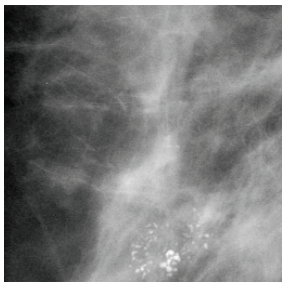
The groups are built from the membership matrix $U = [\mu_{ik}]$ associating the pixel z_k to the group i which has the maximum membership degree. Fig. 1(b) shows the identified groups with U for the drop of milk image. Fig. 1(c) shows two groups and two sub-groups when the typicality values are used. In a more classical approach, Fig. 1(d) contains four groups resulting from the segmentation of the image using the U data.

The criterion to assign a pixel z_k to a group i is based on the maximum membership degree of the pixel to the groups. However, as we only have two classes and the sum of the membership degrees must be one, we can use 0.5 as the threshold to establish the limits of the groups. Fig. 1(b) shows the results with the drop of milk image, where the white color represents the drop of milk, and the black color the background of the image. This result is similar to that obtained with the FCM algorithm. It represents a good segmentation between the drop of milk and the background, except that region in the lower part of the drop of milk, a consequence of its shadow, because it has a gray intensity level close to the value of the prototype (49.024) for the background.

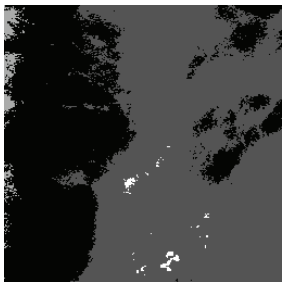
Additionally, the typicality matrix $T = [t_{ik}]$ is used to find the sub-groups into each group. This is done by establishing a threshold α for the typicality values t_{ik} such that each group is divided into typical and atypical data. Data with typicality values above the threshold α are considered typical, and data with typicality values below α are considered as atypical and they are the elements of the sub-groups. The first step is to identify the group of each pixel and then decide which points belong to a group and which to a sub-group.

Fig. 1(c) shows the results of segmentation and sub-segmentation of the image with the typicality values t_{ik} and a threshold $\alpha = 0.2$, according to the approach previously described. Each group and sub-group is represented by different color; dark gray and black represent the drop of milk and background respectively, and white and light gray represent the corresponding sub-groups. Comparing the pixels of a sub-group with the elements of its corresponding group, results in a set of elements (the furthest elements from the prototype belonging to this group) that clearly differs from the most representative elements of the corresponding group.

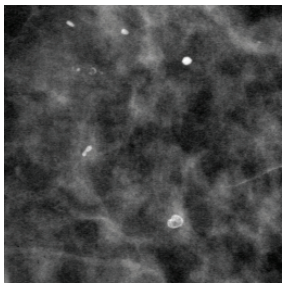
This allows for detection of pixels belonging to the drop of milk and pixels with a marked difference result, in this example, in the illumination. Fig. 1(d) shows the results of the segmentation with the PFCM using four groups and the membership matrix U only. A comparison between Fig.



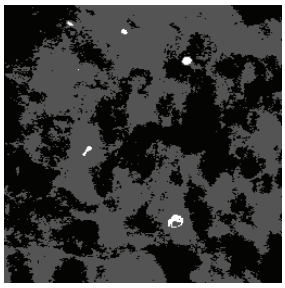
(a)



(b)



(c)



(d)

Fig. 2. (a) ROI image of a mammogram A, (b) two groups and two sub-groups of mammogram A using the matrix *T*, (c) ROI image of mammogram B, and (d) two groups and two sub-groups of mammogram B using the matrix *T*.

1(c) and Fig. 1(d) shows the different results according to the selected approach, and the values of the approach proposed in this work.

Usually, the most important regions in a mammogram are those pixels with a marqued difference in the intensity level with the regions considered as normal. However, the pixels of most interest are the brighter ones. These pixels can be microcalcifications, which are not always so easy to identify, and they are the regions the experts are looking for such that the microcalcifications could be classified as benign or malign.

Fig. 2(a) and Fig. 2(c) show two mammogram images. Fig. 2(b) and Fig. 2(d) contain the results when the PFCM algorithm is applied to identify the groups and sub-groups. Fig. 1(b) shows the results when the threshold is $\alpha = 0.2$ and the PFCM is applied to the mammogram A. In this case, the sub-groups represent points of much interest for detection of breast cancer, because these tissues are less representative of healthy tissue and they can correspond to microcalcifications.

For the results of Fig. 2(d) the selected threshold was $\alpha = 0.05$ because microcalcifications of the mammogram with region of interest (ROI) image show a marked difference and they can easily be distinguished from the healthy tissue. That is the reason why their typicality values t_{ik} are so small. So, the α threshold plays an important role for the image sub-segmentation, and it depends on the gray intensity variations on the image.

Unlike in the drop of milk image, in this case the sub-groups can be used as an aid to medical diagnosis. However, both examples show the practical utility of atypical data identification in segmented images.

IV. Conclusions

In this work we have proposed to use the PFCM clustering algorithm for image segmentation and furthermore to find sub-groups that allows us to obtain more information that aid in diagnosis (for example). In one of the two kinds of images used, the drop of milk, it was possible to identify some imperfections corresponding in this case to effects of illumination. For the other kind of image, the mammography, the sub-groups represent tissue with characteristics that differ from normal tissue, and this could be beneficial during the cancer risk analysis, as these are the kind of characteristics experts look for. Thus, for applications such as mammograms, atypical data could be of greater interest than typical data of the same group. In a forthcoming work we are going to study the impact of the PFCM parameters on the results.

Acknowledgments

This research has been partially supported by University of Guadalajara under project 76817, as well as the UPM, National (MICINN) and Madrid (CAM) Spanish institutions under the following projects: AL09-P(I+D)-12, PTFNN (MCINN ref: AGL2006-12689/AGR).

References

[1] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.
[2] X. J. Fu and L. P. Wang, "Data dimensionality reduction with application to simplifying rbf network structure and improving classification performance," *IEEE Trans. System, Man, Cybern, Part B: Cybernetics*, vol. 33, pp. 399–409, 2003.
[3] N. R. Pal, S. K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilitic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, 2005.
[4] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*, P. Press, Ed. Plenum Press, New York, 1981.
[5] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *International Conference on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.
[6] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, pp. 32–57, 1973.
[7] N. R. Pal, S. K. Pal, J. M. Keller, and J. C. Bezdek, "A new hybrid c-means clustering model." in *Proc. of the IEEE Int. Conf. on Fuzzy Systems, FUZZ-IEEEŠ04", I. Press, Ed.*, 2004.
[8] F. Höppener, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis, Methods for classification, data analysis and image recognition*, Wiley and Son, Eds. Chistester, United Kingdom, 2000.
[9] H.Timm, C. Borgelt, C. Döring, and R. Kruse, "Fuzzy cluster analysis with cluster repulsion." in *presented at the Euro. Symp. Intelligent Technologies (EUNITE), Tenerife, Spain*, 2001.
[10] H. Timm and R. Kruse., "A modification to improve possibilistic fuzzy cluster analysis." in *Conference Fuzzy Systems, FUZZ-IEEE, Honolulu, HI, USA*, 2002.
[11] H. Timm, C. Borgelt, C. Döring, and R. Kruse, "An extension to possibilistic fuzzy cluster analysis," *Fuzzy Sets and systems*, vol. 147, no 1, pp. 3–16, 2004.
[12] N. R. Pal, S. K., and J. C. Bezdek, "A mixed c-means clustering model." in *IEEE International Conference on Fuzzy Systems, Spain*, pp. 11-21, 1997.