Intelligent System for the Automated Diagnosis of Histological Images

J. I. ESTÉVEZ, J. F. SIGUT, J. L. SÁNCHEZ, J. D. PIÑEIRO, S. ALAYÓN, R. L. MARICHAL, J. M. TORRES
Department of Fund. and Applied Physics, Electronics and Systems
University of La Laguna
Edf. Física y Matemáticas, Av. Fco. Sánchez S/N, La Laguna, Tenerife
SPAIN

Abstract: - The problem of diagnosing histological/cytological images is the main objective of this work. A system for the analysis of this kind of digital images is proposed. This system aims to systematize the image processing/machine vision/classification procedures needed to reproduce the physician interpretive abilities and collect them in standardized protocols both reproducible and of quantifiable performance. Flexible soft–computing techniques will be needed to reach an adequate performance at the level of image interpretation, while a knowledge-centric, object-oriented implementation approach will be preferable to hide the details of these procedures from the specialist. A case study of pathology diagnosis is proposed as illustration of the difficulties and the techniques that will be used in the system.

Key-Words: - Image Interpretation, Histological Images, Medical Protocols, Texture Analysis, Structural Image Analysis, Follicular Lymphoma

1 Introduction

The aim of this paper is to outline the development of an image-interpretation prototype system which makes easier creating and spreading protocols for the automated analysis of digital images in order to detect possible pathologies. These images are obtained from light microscopy of histological and cytological samples. We expect to obtain good results by combining conventional and well-documented techniques with some novel approaches using soft-computing techniques.

An important part of the project deals with the systematization of the built procedures. We will try to improve the reproducibility and spreading of protocols for automatic classification of samples. The lack of formalization has been traditionally a serious disadvantage to the use of these protocols in spite of the availability of many useful tools. We expect that integrated tools and techniques can be evaluated and criticized by the same experts that gave the diagnosis guidelines.

Another part of the project deals with the research on combination of image analysis and processing techniques with soft computing methods (neural networks, fuzzy systems, genetic algorithms ...). In this respect, for example, one research line will focus on the problem of texture analysis by means of recurrent systems able to discriminate between series of data. In studies carried out by several groups, including us, this approach has been proven to be successful when applied to the discrimination of cell nuclei based on the chromatin distribution estimated from grey levels. We intend to extend these techniques to the characterization of tissues. In the rest of the paper, we describe the organization of the system and present a first problem that will serve as a test-bench for it.

2 System Description

Several attempts are found in the literature to formalize and standardize medical protocols and guidelines. Systems as Proforma [1] and GLIF [2] are examples of this, among many others. In some of them, the emphasis is put not only in diagnosis but also in monitoring treatments. In our approach, we intend to cover only the diagnosis problem and the focus is slightly different. What is attempted is the automated emulation of already standardized procedures carried out by experts in interpreting this kind of medical images. In a typical scenario of use, the specialist defines a new protocol for a specific pathology by combining available basic building blocks that represent tasks such image processing algorithms (image enhancers, shape and color detectors, object outlining), classification procedures, etc. Frequently, this must be done over a sample case image, in an exploratory fashion until the results are adequate in the specialist criteria. Traceability is other desirable feature in this scenario. The expert must be able to visually reproduce the application of a protocol over a set of images to detect which component of the protocol is not performing well.

The basic building blocks can be grouped in higher level components in a hierarchical fashion, being the full protocols at the top of this hierarchy.

The system is being developed in a rapid-prototyping fashion to allow for the expert physicians quick feedback about its performance. For this purpose, we have chosen Python, which is a popular, object-oriented scripting language. Several modules provide an object-oriented database framework to give persistence to application data. This way, it is not necessary to design a conventional database and an object system model separately. The development is based heavily in incorporating procedures and algorithms able to help in particular image diagnosis problems, as the next case study illustrates.

3 Case Study: Follicle Identification

As a first test case, we have tackled the problem of characterization of a histological feature. This problem will allow us to build the basic system blocks that will constitute the first system prototype.

In many cases of lymphoma, structures known as follicles must be identified to produce an adequate diagnosis. Characteristics of this structure determine the kind and malignity of a possible cancer. Here, we consider two alternative techniques for this purpose.

4.1 Structural Analysis

Structural analysis of images is a complex task where automatic procedures have to include a lot of knowledge coming from areas of expertise very far away from the low level processing aspects: expected appearance of the objects, relevance of the detected elements in the image, lack of information, etc. Therefore, these algorithms contain a lot of free parameters and contextual decisions that are difficult to describe even for experts of the image processing side. Using this kind of software by medical experts is a challenging task and researching how to produce robust algorithms and guidelines for using them in the area of medical diagnosis is one of the main objectives in this project.

Other important aspect that we are considering in this research project is the application of artificial intelligence techniques. There are at least two open research lines here. In first place, we want to find innovative applications of artificial intelligence techniques for image processing. In second place, as was stated above, these techniques should be robust enough to be included in a firmly stated medical protocol.

In order to exemplify the problems that we are finding, the next paragraphs describe a practical problem that we are working with these ideas in mind. Figure 1 shows a sample with one follicle. The objective is to extract those parts of the image that are relevant in deciding about the presence of a follicle. In the figure, we can see how the borders of the structure present a higher density of dark pixels. However, there are other parts in the image with a higher density also. Information about the geometrical aspect of the structure is highly important in this case if we want to differentiate the relevant areas of the image.

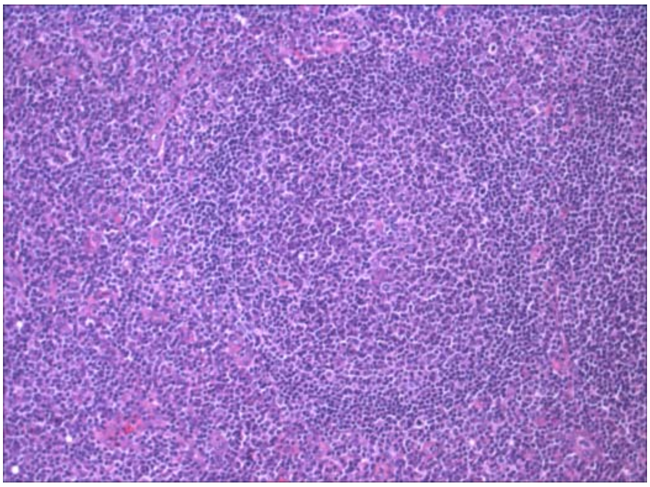

Figure 1. Original follicle image.

Figure 2 shows a grayscale image of the same sample. A preprocessing stage to maximize differentiation between darker and lighter pixels was carried out. This processing stage does not need initialization of image dependent parameters. We can consider this, a robust part of the process. However, this stage of the processing is still far away from a structural analysis.

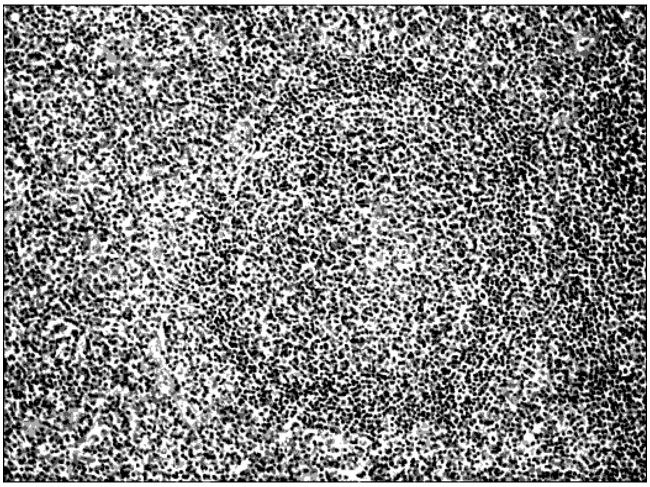

Figure 2. Grayscale image showing a follicle.

In order to extract the main spatial structures a spatial clustering algorithm based on local density maxima is applied. The algorithm uses basin spanning trees to characterize the neighborhood of the local density maxima. The algorithm was proposed in [3] as a clustering by local maxima method with low computational complexity. It has the following steps:

- For each point in the image find the topological neighbors using Delaunay triangulations.
- Compute a density estimate at each point.
- For each data point, find the neighbor with the largest increase in estimated local density. Create a directed edge from the point to this neighbor. If none of the neighbors have a higher density the point corresponds to a local density maximum and it will be considered as root of a tree.
- Each cluster is obtained following the tree edges from the root toward the leaves.

At this stage of the processing an important parameter related with the spatial scale of the image has to be adjusted. This parameter establishes the size of the box used to estimate the local density. The spatial scale considered in the processing of the image is critical because the structure that we are trying to extract has its own scale. Processing in a different scale will cause the loss of the relevant information about the object. Figure 3 shows the result after applying the algorithm.

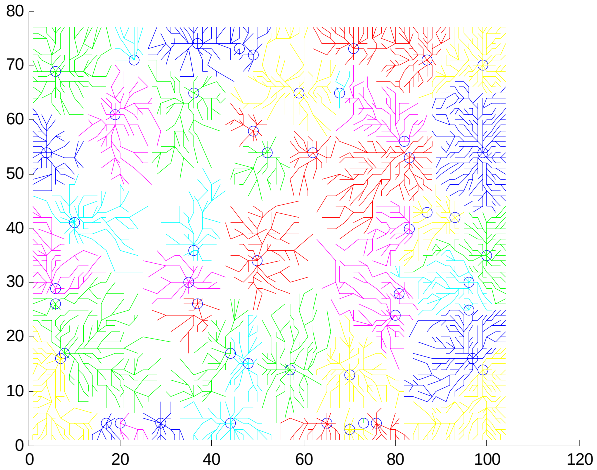

Figure 3. Spanning trees obtained after application of the algorithm based on clustering by local maxima.

It is clear that the obtained trees do not capture the band shaped structure that surround the follicle. This is because the spatial scale of the algorithm was adjusted for the width of this band. However, the band has a much larger length and many local maxima are found through the border of the follicle. For this reason it is necessary to complete the algorithm with another step.

This step tries to merge several trees in a new tree in order to recover parts of the follicle border. The fusion process is based on two features: similar density at the local maxima and small distance between the trees and between their roots. This algorithm has to set three parameters that control how the trees are merged. The first one is for the density while the other two are thresholds for minimum distances between the trees and between the roots. Figure 4 shows the results. Several trees contain now part of the band shaped border of the follicle.

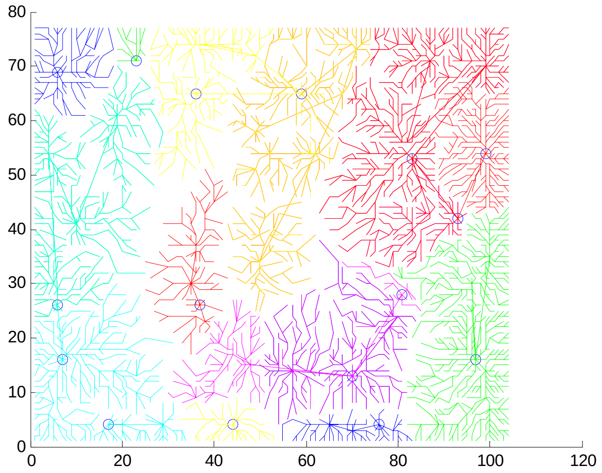

Figure 4. Tree fusion based on both density and distance measurements.

At this example we see how the number of free parameters and in general the difficulties to produce a systematic protocol increase when our objectives move from the low level to the high level processing.

From the point of view of the intelligent systems, structural image analysis produces results where the application of specific techniques for automatic algorithm production is at least an interesting research area. Genetic programming [4] is a methodology that has been used for automatic production of algorithms, data mining and knowledge discovery. We are researching its application for image processing in two directions. The first one has been already proposed in [5] and it consists of the automatic production of mathematical algebraic expression combining the results of different features (for example, texture features) to obtain a good classifier.

The second one deals with the production of “structural filters”. To understand this methodology, we can propose the following problem: improve the algorithm described for the previous example with a method to quantify the degree of membership of a tree to the class of the ones that typically are found in the border of a follicle. It is clear that other techniques based on geometrical features could be used in this case. However, we follow this approach to describe the type of structural analysis application where a genetic programming system can be used.

The first step is to encode the tree in a chain of characters describing relevant geometrical features. In our case, the characters could describe the angle of each child relative to its parent at each node of the tree. This is an example:

(ACEF(HH(A)A(G))G(B(B(A(H)C(HC(B(D(D)C)B(A(DC)BB)D))))B(CE)D(A(B(H)))D(C(C(CE))C(B(C(D))C(D)))C(EE)))

The second step is to establish a grammar where the evolved program is described. The kinds of programs that we are testing for this application are regular expressions. A regular expression is used to describe a pattern of characters. In the genetic programming framework, the regular expressions are individuals of a population that is evolved. A fitness measure is computed for each individual. This measurement is based on how differently the regular expression matches in trees form each class. Using regular expressions as evolved programs has been previously researched in the context of bioinformatics applications [6]. The first stage of research on the example described above is being carried out with the PerlGP system [7].

It is clear that this application is a machine learning problem that needs the construction of the training and test sets. The correct construction of the training and testing sets is one of the tasks where the proposed environment could be used to supervise possible common mistakes as unbalanced proportion of members of each class, or introduction of some bias. For example, if we tend to choose follicle trees for the training set that are shorter than the ones selected form the rest of the image we will obtain an evolved program where the regular expressions obtained will be oriented to count the number of characters of the string representing the tree and not possible hidden structures (Figure 5).

```
sub evaluateOutput {
  my ($self, $data) = @_;
  my ($x, $y, $z, @output);
  my $temp;
  foreach $input (@$data) {
    $x = $input->{x};
    $temp=$x;
    $y=()= $temp =~ /\([ABCDEFGH]+\)/g;
    # end evolved bit
    push @output, { 'y'=> $y };
  }
  return \@output;
}
```

Figure 5. Evolved code obtained with the PerlGP system with a bad designed training set.

4.2 Statistical Texture Analysis

Texture analysis plays an important role in many image analysis applications. In particular, texture methods have been extensively used in medical image analysis [8], [9]. One of the usual ways to deal with textures is by considering them as random phenomena. In this context, the formation of a texture is described with the statistical properties of the intensities and positions of pixels. The co-occurrence probabilities provide a second-order method for generating texture features [10]. These probabilities represent the conditional joint probabilities of all pair wise combinations of gray levels in the spatial window of interest given two parameters: interpixel distance and orientation. The probabilities are stored in a sparse matrix referred to as the gray level co-occurrence matrix or GLCM. Many different statistics can be derived from this matrix. The ones used in this work are the following:

Uniformity (U): $\sum_{i,j=0}^{G} C_{ij}{}^{2}$

Entropy (E): $-\sum_{i,j=0}^{G} C_{ij} \;\log C_{ij}$

Dissimilarity (Di): $\sum_{i,j=0}^{G} C_{ij} \;|i-j|$

Contrast (C): $\sum_{i,j=0}^{G} C_{ij} \;(i-j)^2$

Max. probability (M): $\max\{C_{ij}\}$

Homogeneity (H): $\sum_{i,j=0}^{G} \frac{C_{ij}}{1+|i-j|}$

Directivity (Dr): $\sum_{i=j} C_{ij}$

where C_{ij} represents the co-occurrence probability of gray levels i and j.

Before proceeding with the calculation of these statistics, the different parameters involved in the problem must be set. For the purpose of this preliminary study, the following values were chosen for them:

- Window size and shape: square window with a side length of 8 pixels.
- Number of gray levels: 32.
- Interpixel distance: 1 pixel.
- Orientation: 90 degrees.

A training set of 15 follicle and 15 non-follicle samples was selected from one image in our database. These samples were used to build a Bayesian classifier assuming normal distributions for both classes and equal prior probabilities. Due to the small number of samples, just two statistics were considered as input features to avoid dimensionality problems. All pair combinations were tried and the best results were obtained with (Di,H) and (E,H). Figs. 6, 7, 8 and 9 show the classifier performance on two gray level 1550*2088 images containing follicle and non-follicle areas. In both cases, some follicle areas can be identified because of the limited presence of white pixels with respect to the surrounding regions. It is clear that the results of these first trials are not too satisfactory but it is expected that a more exhaustive analysis can provide better results in a near future.

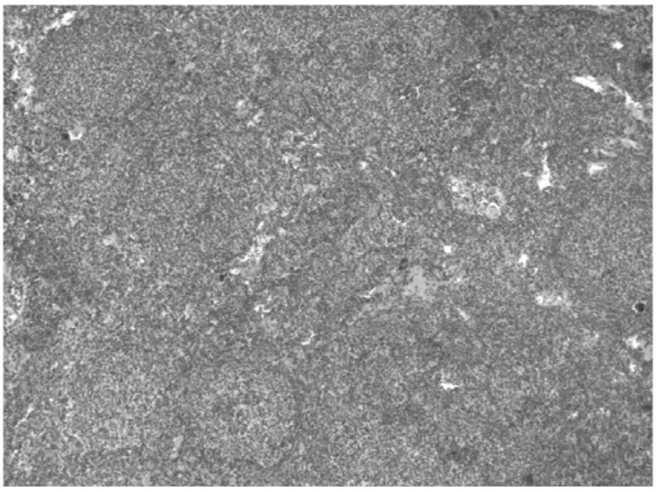

Figure 6. First original test image

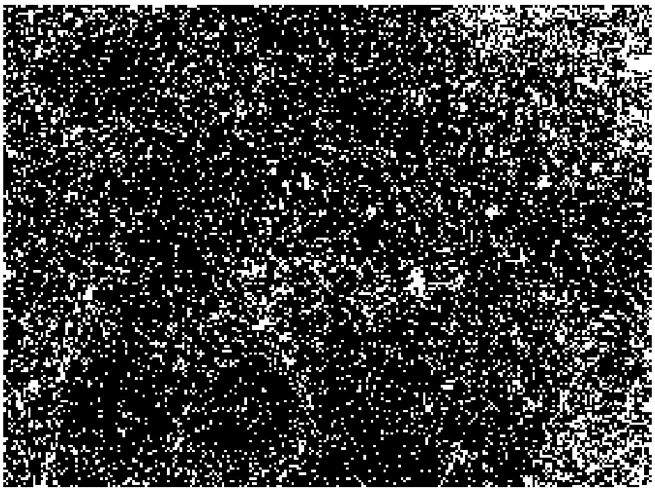

Figure 7. Binary image obtained from figure 1 after the classification

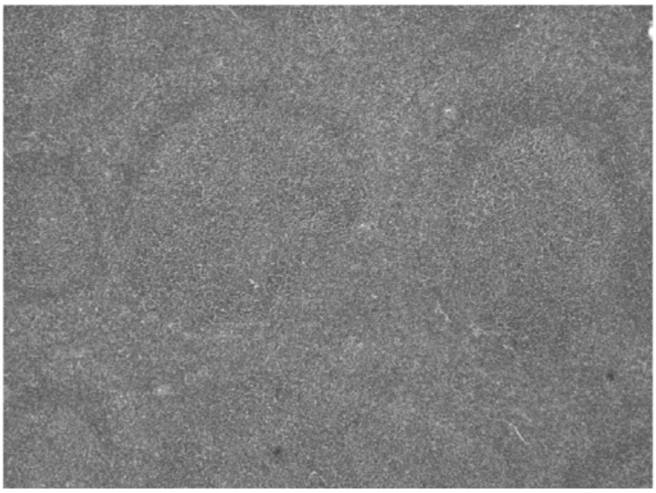

Figure 8. Second original test image

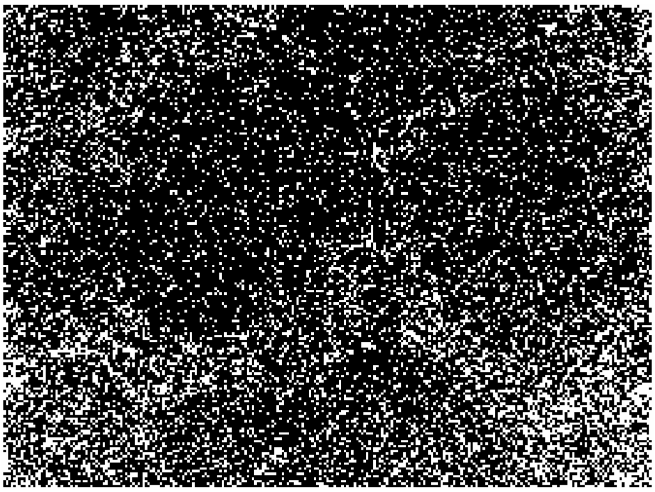

Figure 9. Binary image obtained from figure 3 after the classification

4 Conclusion

A work in progress about a system for automating the complex task of diagnosing histological images was presented. Several lines of development are being pursued simultaneously, from application usability issues, to novel soft-computing techniques of image analysis. The system is still in an early phase of development and we aim for the first prototype, integrating the most basic functionalities but with a great deal of the final structure built-in, will be ready as a usable interactive tool to create/critique/refine the integrated procedures and protocols by the pathologists without the help of the application designers.

Acknowlegments:

This work is supported by the Spanish *Ministerio de Ciencia y Tecnología* and the European *Fondo Europeo de Desarrollo Regional (FEDER)* with the research project grant number TIC2003-05046

References:

[1] J. Fox, N. Johns, A. Rahmanzadeh, R. Thomson "Disseminating Medical Knowledge: The PROforma Approach" *AI Med.* 1998; 14:157-81

[2] L. Ohno-Machado, J. H. Gennari, S. Murphy, N. L. Jain, S. W. Tu, D. E. Oliver et al. "The Guideline Interchange Format: A Model for Representing Guidelines", JAMIA 1998. 5(4): 357-72

[3] Sören Hader y Fred A. Hamorecht. "Efficient density clustering using basin spanning trees". *Proceedings of the GfKl 2002,* Springer. 2003.

[4]. J. R. Koza. *Genetic Programming: On the Programming of Computers by Natural Selection*. MIT Press, Cambridge, M.A. 1992.

[5] J. M. Daida, J. D. Hommes, T.F. Bersano-Begey, S.J. Ross, J.F. Vesecky. "Algorithm Discovery Using the Genetic Programming Paradigm: Extracting Low-Contrast Curvilinear Features from SAR Images of Artic Ice". *Advances in Genetic Programming II. P. Angeline and K. Kinnear (ed.)* The MIT Press. 1996

[6] A Heddad, M Brameier, R. M. MacCallum. "Evolving Regular Expression-based Sequence Classifiers for Protein Nuclear Localisation". *2nd European Workshop on Evolutionary Bioinformatics*. 2004

[7] R. MacCallum. *Perl Genetic Programming environment (PerlGP)*: http://perlgp.org/

[8] S. Petroudi and M. Brady. "Classification of Mammographic texture patterns". *International Workshop on Digital Mammography,* 2004

[9] Y. L. Huang, K. L. Wang, D. R. Chen and Y. K. Liu. "Diagnosis of Breast Tumors with Ultrasonic Texture Analysis Using Support Vector Machines". *Neural Computing and Applications, (SCI, EI)*, 2004.

[10] R. Haralick, K. Shanmugan, I. Dinstein. "Textural features for image classification". *IEEE Transactions on System, Man and Cybernetics,* 3 (1973) 610-621.