Toward Abnormal Nuclei Detection Using an Integrated Automatic System

Ramona GALATUS, Sorina PERSA, Daniel MOGA, Viorel TRIFA,
Liliana NEAGA, Tiberiu MARITA and Radu MUNTEANU jr.

Abstract: *The traditional process for detecting the cervical cancer is called Pap smear testing and it is the most widely used screening technique. The pathologists diagnose the smear according to its normality or abnormality. The huge number of slides to be analyzed requires an automated computer-aided system which can help in diagnosis process. This paper proposes an architecture appropriate for a system that automatically scans the slides and extracts the regions of interests looking for signs of precancerous and cancerous changes, and categorize them according to a series of criteria.*

1. Introduction

The traditional process for detecting the cervical cancer is called Pap smear testing and it is the most widely used screening technique. In the examination process the cells are collected from the uterine cervix and, after an initial preprocessing, are put on a glass slide and are sent to the cytology laboratory to be examined under the microscope for signs of precancerous and cancerous changes. The pathologist will diagnose the smear according to its normality or abnormality. Based on the physician's decision, the subject will follow colcoscopy, biopsy and treatment.

The reason why the detection of cervical cancer is crucial is that, if detected early, it has an almost 100% chance of cure. However statistics have shown that many women still died of cervical cancer. The high mortality rate has two major reasons: the first one is the fact that the women are not following the screening procedure and the second one is that some of the slides are diagnosed as false negatives.

Usually, women of 20 up to 60 years old are following this screening procedure (over 50 millions women per year in America), so the number of slides that must be examined is huge. Thus, the time allocated to the investigation of a slide is very short (about 10-15 minutes). There are some fundamental characteristics that influence the investigation process:

– The huge number of slides with no cancerous signs that are being analyzed.
– The huge number of cells examined on each slide
– The big number of images (slide's zones) that must be explored.

Statistics prove that over 90% of the slides examined by a pathologist are normal, and each slide may contain up to several tens of thousands of cells. Beside the fact that the slides having tumor are very rare, they have approximately only 100-200 suspect cells on the whole surface of the slide. That is the reason why it is very difficult, even impossible for the human eye to detect all cases of early cancer. Due to large volume and fast turnover of the test slides in a typical histology laboratory, when relying only on manual inspection, it becomes inevitable that some abnormal Pap smears will be missed, meaning that some positive test results are sometimes overlooked. Even the best laboratories can miss from 10% up to 30%

abnormal cases. When this occurs, a potentially curable carcinoma-in-situ can progress to an invasive cancer.

A continuous range of abnormal nuclear morphology is seen in epithelial cells in cervical smears [2]. The morphology reflects those abnormalities of the cervical epithelium which involve the cells on the surface. Hence a simple basal cell hyperplasia does not produce changes at the surface of the epithelium or in the cervical smear. Lesser changes in the cells in the smear on their own are normally associated with inflammatory or reactive conditions, which are essentially benign. The more strikingly abnormal features described as dyskariotic, as recommended and defined in 1986 by the Working Party of the British Society for Clinical Cytology, are associated with cervical intraepithelial neoplasia (CIN). Cervical intraepithelial neoplasia (CIN) is characterized by some distinct features [2]: *disproportionate nuclear enlargement, which leads to high nuclear-to-cytoplasm ratio, hyperchromasia, irregularity in form and outline of the nuclei, irregular chromatin distribution, presence of keratinization, abnormalities of the number, size and form of the nucleoli, multinucleation.* In the case of cervical intraepithelial neoplasia three degrees of evolution have been identified: 1 (mild), 2 (moderate) and 3 (severe). The features of each grade differ, although, in a smear, cells belonging to different grades can be encountered [2].

2. Overview of the techniques appropriate for automatic systems

The above mentioned facts are clearly demonstrating that is worth to investigate the architectures appropriate for the design and implementation of a system able to automatically handle the problem of recognizing abnormal nuclei in a series of slides.

In the literature are found more attempts to realize such a system, more or less finalized. A working one, used in USA is PAPNET [1] system, which is an artificial neural networks-based screening method. PAPNET uses two separate neural networks: one is trained to recognize suspicious single cells and the other to recognize abnormal cell clusters. In a test of 1247 Pap smears, the PAPNET system identified 517 of 534 abnormal smears, which represents a 97% accuracy rate. The PAPNET is a supplementary screening method that eliminates the searching process, reduces human fatigue and improves accuracy of diagnosis leaving in the responsibility of a qualified pathologist to make the final diagnosis.

Another approach based on fractal geometry was used to characterize irregularly shaped and complex figures, and as a result, to identify the abnormal cells [3]. The fractal dimension computed using a box-counting algorithm was chosen to characterize the structures shape. The results that have been obtained proved that the abnormal nuclei do possess a distinct fractal dimension compared to that of the typical nuclei [2]. In spite of that, the fractal dimension of atypical nuclei in CIN lesions alone is not enough for a histopathological diagnosis and other criteria must be used to characterize the grade of CIN: size of nuclei, the pleomorphism, the staining intensity, the pattern of chromatin, the appearance of nucleoli and/or vacuoles, mitotic figures.

Another attempt in the automated detection of CIN was done at the University of Colorado at Boulder, using optics [4]. The process was divided in two stages. First slides containing abnormal cells (enlarged nuclei) are detected, using an optical implementation of the hit-miss algorithm [8] using an optical correlator, which compares the Fourier components of an input image (the image of the microscope slide) with that of the template (shape) of interest. A more complex, digitally implemented, and time-consuming second stage analysis is then performed on the remaining suspicious slides to characterize the degree of CIN.

Another method in development, called direct visual inspection (DVI), uses light to detect cellular changes in cervical tissues [6].

In this case, no sample is taken. Instead, light is shined into a woman's cervix. The way in which light reflects back gives physicians a full image of the cervix, allowing them to mark exactly the location of precancerous and cancerous conditions. Essentially, the light detects structural changes in the cells of the cervical tissue without removing, testing or treating any of those cells from the cervix. The main drawbacks of this method are the facts that its specificity is very low and there is a very high level of user variability in results due to the technique and experience in using the instrument. Also, the training is complex and the procedure is expensive.

In the present paper we propose a system which could perform the automated screening of the cytological slides and the first phase would be the automatic detection of the slide regions containing abnormal cells. The main dyskariotic features of the cells which can be used for this processing phase is the disproportionate nuclear enlargement, which leads to high nuclear-to-cytoplasm ratio. The slide regions containing such cells will be retained for further automated or analysis or final diagnosis by qualified pathologists.

3. Overview of the proposed architecture

The functionality of the proposed architecture is based on the use of algorithms for automated image analysis applied to series of images of the cytological slides, obtained using a slide positioning system and an image acquisition system. The purpose of these algorithms is the selection of the image regions which are contained in tissue regions, cells or other cellular constituents with modified properties, which are suggesting abnormality. The result of their applying will be image series containing the suspect regions of the slide and the associated qualitative and quantitative indicators. These image series will be examined by a specialist for final diagnostic, either on a local workstation or from a remote workstation through the telemedicine application. The automated scanning process will be done using two systems, based on intelligent modules with DSP and microcontrollers: a manipulating robot for supplying the microscope with slides and a positioning system for planar movement of the slides under the microscope. A database can be developed for storing the relevant images along with the associated diagnostics. Along with the software packages for acquisition, processing, storage, encrypting and communication of the microscopic images, that kind of database can be used for the training of the resident physicians, master and PhD students.

4. Automated detection of abnormal nuclei

From the image processing point of view, the detection and classification process of the Cervical Intraepithelial Neoplasia (CIN) can be divided in two main phases:

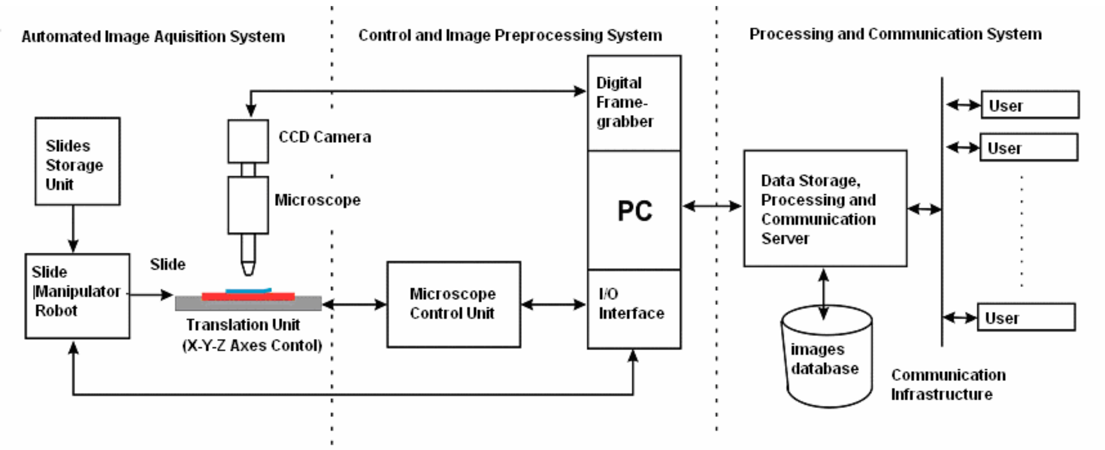

Automated Image Aquisition System
Control and Image Preprocessing System
Processing and Communication System
Slides Storage Unit
Slide Manipulator Robot
Slide
CCD Camera
Microscope
Translation Unit (X-Y-Z Axes Contol)
Microscope Control Unit
Digital Frame-grabber
PC
I/O Interface
Data Storage, Processing and Communication Server
images database
User
User
User
Communication Infrastructure

Fig. 1. Schematic View of the proposed architecture.

- preprocessing phase to detect slide regions with abnormal nuclei
- classification of the CIN level in the three main categories: CIN1, CIN2 and CIN3.

The main focus in this paper will be on this first phase: the detection of slide regions with abnormal nuclei. The proposed approach is based on two steps (Fig. 2): the segmentation of nuclei and detection of the abnormal nuclei.

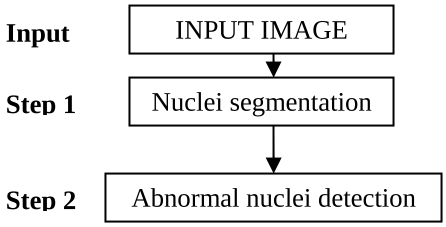


Fig. 2. Steps for the preprocessing phase

4.1. Nuclei segmentation

Having an image with multiple cells, each cell having one or more nuclei (in the case when multinucleation is present), we want to obtain a binary image where only the nuclei are present. This step is one of high difficulty and a very important one, because a wrong segmentation of the nuclei would lead to a false detection, which turns into a false diagnosis. There are several general approaches which seem appropriate for the segmentation task:

- Image thresholding
- Color-based segmentation (clustering)
- Edge detection-based segmentation

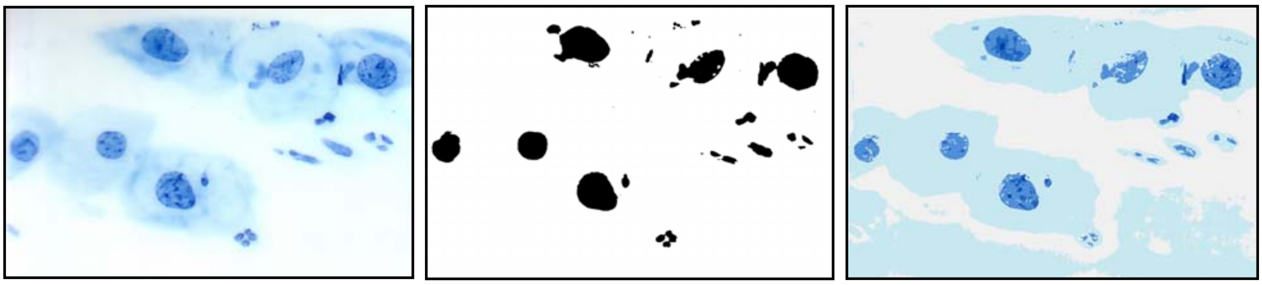
Thresholding is the most simple segmentation method when objects with well defined color or intensity are separated from the background using image binarization. For the case of PAP smear images the colors present in the nuclei and in the cytoplasm are very much alike, especially when referring to abnormal nuclei, which possess hyperchromasia, thus the nuclei wont be easily distinguishable from the cytoplasm and further processing would be very difficult. A result of the binarization process can be seen in the figure 3.

Methods based on color segmentation as the one presented in [7], based on color clustering in SCT-center and RGB space are sensitive to the tissue and especially to the stain color of the slides and thus are not appropriate for a fully automated system. We are interested in implementing a fully unsupervised method. The results obtained by the segmentation of the same input image (fig. 3.a) with the method presented in [G7] with an RGB cluster dimension of 49 and the number of 16 SCT-space colors is shown in figure 3.c. As it can be seen the cytoplasm was not correctly segmented and too many color clusters have remained in the nuclei, due to hyperchromasia of abnormal cells.

Edge detection-based segmentation

The segmentation method chosen in our approach is based on edge detection. The main reasons for choosing are good segmentation results and no need of supervising from the user. Its steps are presented in the following diagram.

The first step in the nuclei segmentation process is the classical transformation of a RGB image to a grayscale one, and was compulsory for the applying the edge detection method. The second step is the edge



a. b. c.

Fig. 3. **a.** Original image; **b.** Binarized image ; **c.** Resulting image after color based segmentation.

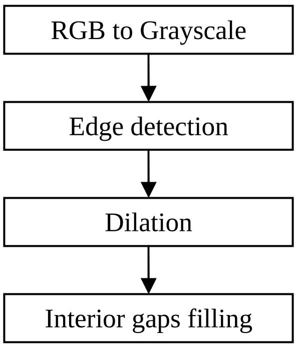


Fig. 4. Nuclei segmentation steps.

detection, which was performed using a gradient based method implemented with the Sobel mask [8].

Applying a simple Sobel mask is not enough in order to obtain thin (one pixel wide) edge lines (fig.5.a). Because edges between nuclei and cytoplasm are not well defined, classical edge thinning techniques as non-maxima suppression or hysteresis thresholding would fragment the nuclei contour. That is the reason for implementing the following algorithm as a better way to obtain one pixel wide edge lines. A pixel belongs to a line, only if:

- the gradient amplitude is higher than a threshold;
- either the variation on Ox (horizontal direction) is bigger than the variation on Oy (vertical direction) and the amplitude of the gradient in the chosen point is greater than that of the two neighboring points on the Ox direction or the variation on Oy is bigger than the variation on Ox and the amplitude of the gradient in the chosen point is greater than that of the two neighboring points on the Oy direction.

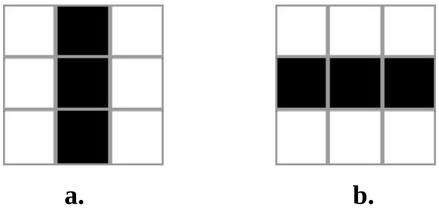


Fig. 5. **a.** Vertical line structural element; **b.** Horizontal line structural element.

Both conditions must be fulfilled in order to obtain a one pixel-wide edge.

In order to be able to fill the nuclei, a further step is required. This consists in two successive dilations applied on the resulted image, first of them with a vertical line-like structuring element (fig. 5.a) and the second of them with a horizontal line-like structuring element (fig. 5.b).

The last step in the nuclei segmentation process is interior gaps filling. This is done using an image reconstruction algorithm. A hole or gap is defined as a set of background pixels that cannot be reached by filling in the background from the edge of the image. A selected point inside the hole where the filing process starts is called marker and the set of all markers is called marker image. The input image is called mask. The binary morphological reconstruction simply extracts the connected components of an image *I* (the mask), by another image *J*, which is a sub-image of the original one (the marker image) [6]. That was implemented in the following way:

- Compute the marker starting from the input image, by taking the complement of a one pixel wide boundary of the input image, while filling its interior with black.

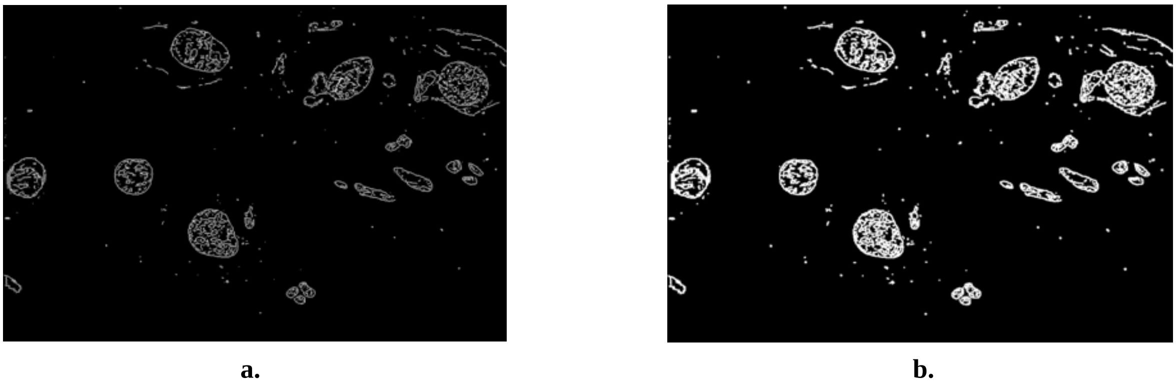


Fig. 6. a. Results after applying the Sobel operator; **b.** Results after applying successive dilations.

- Perform the image reconstruction, using the fast hybrid reconstruction algorithm (which is one of the best algorithms known in the literature and works fine with both binary and grayscale input images) [6].

The result of applying the filling of the holes can be seen in the following figure 7.a. As one can see, in the image above there are six abnormal nuclei. Four of them are segmented correctly, while the other two are not, because the cytoplasm, in the proximity of the nuclei, takes similar colors and, during the edge detection process, the stains from the cytoplasm were identified and were linked to the nuclei. Therefore a last step which eliminates noise and atypical shapes is performed. First, we label the image and then we compute the area of each label in pixels, and eliminate those labels that have an area smaller than a certain threshold. The threshold is set to the minimum area of an abnormal nucleus. Still, some noise elements will be present in the image. These can be eliminated by computing the shape factor of each label (4.1). All the labels that have a shape factor smaller than a specific threshold and an area smaller than another specific threshold will be converted to background.

$$Shape\_factor = \frac{4 \times \pi \times Area}{Perimeter^2} \qquad (4.1)$$

4.2. Abnormal nuclei detection

From the nuclei segmented in the previous step, we want to detect those which are atypical. For that purpose we have used an improved algorithm of the hit-or-miss transformation [8]. The hit-or-miss transform can be used to detect objects of a certain shape. In this way, every circular nucleus with the diameter in a specified range will be marked as abnormal.

Usually, knowing the exact form of the shape that we are looking for, we select the structural element as being of the desired shape, and we choose a local background as the outer border of the structural element, one pixel wide. However, in our case, the abnormal nuclei do not have a standard circular form so we cannot say that all the nuclei are disks with the diameters between a *minimum_diameter* and *maximum_diameter*. The atypical nuclei have irregular shapes, and all we know is that they can be comprised between the minimum and the maximum diameters. This means that a circle having the diameter equal to the minimum diameter will always be contained in an abnormal nucleus, while a circle with the diameter equal with the maximum diameter plus two will always contain any abnormal nucleus. That is why the structural element will be a disk with the diameter equal to the minimum diameter, and the local background will be a circle of diameter equal the maximum diameter plus two, and of width equal to one. This is the only possible choice that will identify the atypical nuclei, but it has a serious drawback: if there are any noise elements or other nuclei around the nucleus that we are trying to identify, at a distance smaller than the ray of the local background, distance measured from

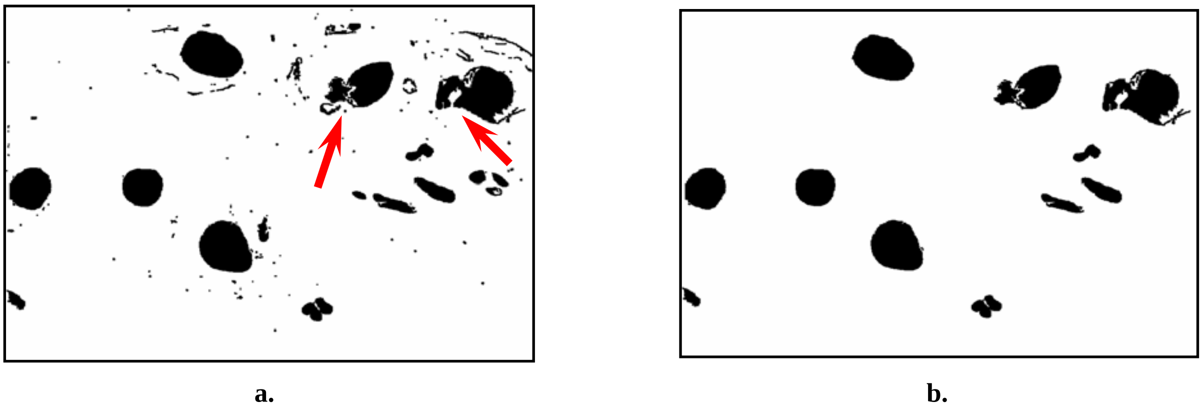


Fig. 7. a. Image reconstruction result; **b.** Image resulted after the removal of the noise.

the origin of the nucleus of interest, the process will fail. The match of the structural element will succeed, but the match of the local background will not. If the dimension of the maximum diameter is big enough, the matching of the local background is a time-consuming operation.

Another problem raised by the local background dimension is encountered when a nucleus is close to the border of the image. If the distance from the center of mass of the nucleus to the border of the image is less than the ray of the local background, that nucleus will not be detected (the match of the structural element will succeed, but that of the local background will fail). In order to solve this problem, we have used the following strategy: if the erosion of the image with the structural element identifies at least one point close to the border of the image while performing the erosion of the image complement with the local background, we will apply only parts of the local background kernel, and act as if a border of width equal to the difference distance between the maximum and the minimum ray and of background color has been added.

Another improvement that can be made in the first step (nuclei segmentation) is to remove the labels with the area bigger than the area of the disk having the maximum diameter, because these labels will not turn into nuclei of interest, but they will definitely fail when trying to match the local background. In the same time, the labels with the area bigger than the threshold mentioned before could represent clusters of nuclei. If no abnormal nucleus has been found in the image, but a label with the area bigger than the threshold has been discovered, the image can be marked, and the cytologist physician can double check this problem.

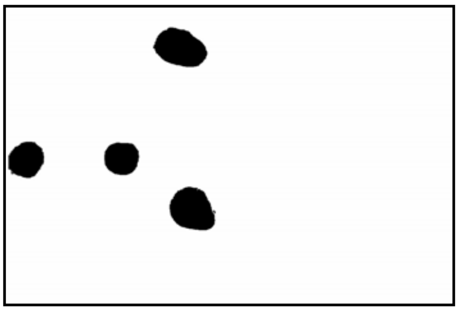


Fig. 8. The result of the nuclei detection by hit-or-miss transform.

The final results of applying the above mentioned techniques for the detection of abnormal nuclei from the considered input image are presented in figure 8.

5. Conclusions and future works

The system architecture proposed in this paper is able to cover most of the functionality expected from an integrated automatic system intended to support diagnosis based on smear screening. While the speed of such system is of primary importance, its smartness, in terms of successfully recognizing and categorizing abnormal areas is even more important. That “smartness” relies entirely on the effectiveness of the processing algorithms. That is the main motivation for which finding appropriate models, able to translate the knowledge of the pathologists into patterns suitable for algorithmic approaches, is in our view, the one who should dictate the roadmap of future research in these topics.

The results of the algorithms presented in this paper are recommending them as a valid solution for the first stages of processing for abnormal nuclei detection. The next stages require a more complex approach, based not only on simple geometric features, but knowledge-based algorithms able to categorize the objects in the regions of interest with an extra accuracy.

References

1. “*Papnet* improves screening efficacity for cervical cancer”, *Technology Application Repor*t, 1997.
2. Winifred Gray, “Diagnostic Cytopathology”, *Churchill Livingston* 1995.
3. R. Sedivy, Ch. Windischberger, K. Svozil, E. Moser, G. Breitenecker, “Fractal Analysis an Objective Method for Identifying Atypical Nuclei in Dysplastic Lesions of the Cervix Uteri”, *Gynecologic Oncology 75*, pp.78-83, 1999.

4. Sunny Bains, “Optical correlators may reduce false negatives in cervical cancer screening”, *OE Reports*, May, 2000.
5. Francis L. Hutchins, Jr., “Progress in cervical cancer detection”, November 2003. http://www.iscpubs.com/articles/abl/b0111hut.pdf.
6. Luc Vincent, “Morphological Grayscale Reconstruction in Image Analysis: Application and Efficient Algorithms”, *IEEE Transactions on Image Processing*, Vol. 2, No. 2, April 1993.
7. D. Gui, T. Marita, R. Galatus, „Area Computation of AgNOR as a Prognostic Marker in Tumoral Pathology”, *Journal of Applied Medical Informatics*, *Vol.8*, *nr 1-2/2001*, pag. 48-54, Cluj-Napoca, Romania.
8. Rafael C. Gonzales, Richard E. Woods, *Digital Image Processing*, 2001.