# Automatic leukemia diagnosis

*D.M.U. Sabino[1]; L.F. Costa[1]; S.L.R. Martins[2], R.T. Calado[3] and M.A. Zago[3]*
*[1]Instituto de Física de São Carlos – dani@if.sc.usp.br; luciano@if.sc.usp.br; [2]Centro de Medicina Diagnóstica Fleury - sergio.martins@fleury.com.br; [3]Centro de Terapia Celular e Departamento de Clínica Médica – FMRP – USP – marazago@usp.br*

## Abstract

Despite the great demand for blood smear analysis in Brazil and worldwide, relatively few efforts have been directed to the automation of this important problem. This paper presents a prototype of a semi-automatic leukemia diagnosis program, emphasizing basic steps in pattern recognition as segmentation, filtering and feature extraction. A supervised learning process segments blood smear images into four regions of interest: nucleus, cytoplasm, erythrocytes and plasma according to their color. Then, the measurements are performed, both general such as perimeter, area, factor form, circularity as well as innovative measures as curvature, skeletons and multiscale fractal dimension, which can provide more objective subsidy for diagnosis.

**Keywords:** Pattern recognition; leukocyte morphology; quantitative microscopy, Bayesian segmentation, multiscale analysis.

## Introduction

Routine diagnosis and classification of hemopoietic elements depends heavily on morphological descriptions of the cells using the optical microscope [1], an exhaustive and repetitive work performed by expert operators. Although this process can be optimized, the available automatic systems do not provide sufficient performance for accurate diagnosis of malignances yet, once most of them are designed to screen for normal cells. This paper describes software for blood smear image analysis, where blood images are split into regions of interest (ROI) by Bayesian color segmentation and analyzed by traditional and innovative features such as the curvature [5,12], skeletons [8] and fractal dimension [7] of the nucleus. The preliminary results allowed characterization of healthy leukocytes, emphasizing the potentiality of the measures for automated diagnosis of abnormalities.

## Materials and Methods

### General structure of the system

We analyzed peripheral blood Leishman stained (Sigma Co.) smears of normal individuals as presented in Table 1. This image collection has different number of leukocytes from each class in virtue of the different concentrations of each cell in the blood. A cytologist captured the images with a Zeiss microscope under 1,000x magnification (1μm≈15pixels) and identified the micrographs morphologically, keeping them in a database.

**Table1.** Number of leukocyte images considered in this article

| eosinophil | neutrophil | basophil | lymphocyte | monocyte |
|---|---|---|---|---|
| 10 | 20 | 5 | 20 | 20 |

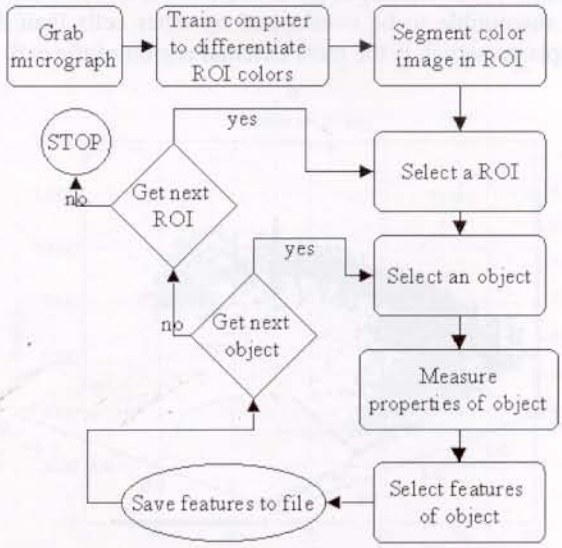


**Figure 1.** Stages in the analysis of a blood smear image

The standard stages in the automated process [17] are illustrated in Figure 1, where the user selects sample areas of a color image (1,000 pixels), labeling them according to the region of interest (ROI) to which they belong

(*computer training*). The training data feed the Bayesian supervised learning algorithm [11] to classify the ROIs (*segmentation*) based on Gaussian probability distribution functions of the color (RGB) pixels, estimated using the training data.

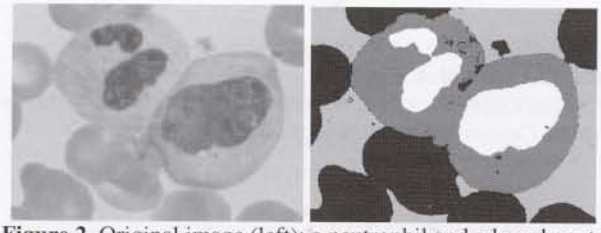

**Figure 2.** Original image (left): a neutrophil and a lymphocyte and its decomposition in ROI (right), after Bayesian color segmentation and morphological filtering.

By filtering the ROI, we divide the image into nucleus, cytoplasm, erythrocyte and plasma (Figure 2). It is necessary to split each ROIs into connected components with pixel area greater than 1,000 (*an object*) because minor pixel areas generally represent noise of the segmentation or isolated granules of granulocytes.

**Measurements**

Once we have separated the objects, several features are extracted from them, including traditional measures such as perimeter, area, nucleus/cytoplasm ratio, circularity, texture, factor form [9], the color of the cytoplasm (Figure 3) and less conventional measures such as curvature, skeletons and fractal dimension. We concentrated on the nucleus measurements, since this is the region where abnormalities often indicate malignance [4,16] and it is less susceptible to be overlapped by other cells than the cytoplasm, which is the most external region of the cell.

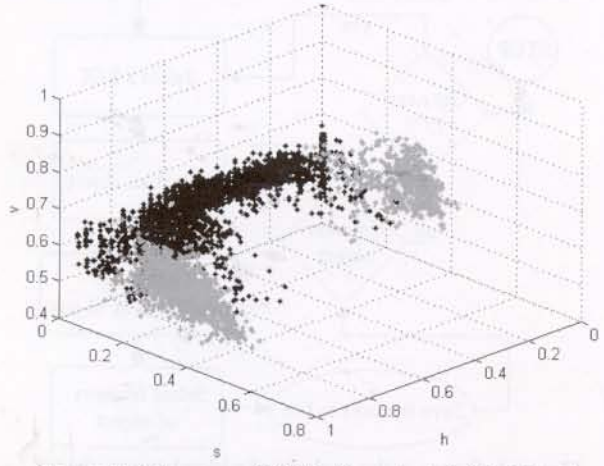


**Figure 3.** Color space (HSV) of neutrophil (black) and eosinophil (gray) cytoplasm: advantageous separability of the leukocytes in terms of HS axis.

Frequently, the nucleus contour of the eosinophil and neutrophil are similar, mainly in the presence of overlapped segments that compose their nuclei. Therefore, an alternative analysis to cytoplasm description was to use its color instead of its contour. Although RGB representation does not allow good separation of the different leukocytes, we verified that they could be differentiated based on the HSV color space, by considering the projection on HS plane as presented in Figure 3. This graph was obtained considering 1,000 pixels from each different image of a set of 6 eosinophil and 6 neutrophil. Such a graph was divided into regions corresponding to the several classes and used as template for identification of eosinophils and neutrophils in other 20 images, allowing 100% correct classifications.

A measure invariant to rotations, translation and reflections of the curve is the curvature. We use signed curvature, where the sign provides indication about the concavity at each contour point. Corners are associated with high absolute curvature values that exceed a threshold. In this article, the contour was captured in anti-clockwise sense, $k>0$ means concavity and $k<0$ means convexity (Figure 4). The curvature ($k$) points over the nucleus contour $(x,y)$ convoluted with Gaussians $g(\sigma,t)$ can be expressed as presented in equation (1):

$$k(t,\sigma)=\frac{x'(t,\sigma)y''(t,\sigma)-x''(t)y'(t,\sigma)}{(x'^2(t,\sigma)+y'^2(t,\sigma))^{3/2}} \qquad (1)$$

where $x(t,\sigma)$ and $y(t,\sigma)$ can be calculated using equation (2) and (3), respectively.

$$x(t,\sigma)=x(t)*g(\sigma) \qquad (2)$$

$$y(t,\sigma)=y(t)*g(\sigma) \qquad (3)$$

The hierarchical descriptor called multiscale curvature [2,3,5,6] expresses the contour curvature in terms of analyzing scales for detecting roughness, corners, as well as curvature statistics such as mean, median, variance, standard deviation, entropy, moments, etc. We calculate curvature using the contour points as one-dimensional complex signal, where the $(x,y)$ coordinates are expressed as complex numbers in the form $(x+iy)$. The derivatives of the signal, calculated for the curvature estimation, uses Fourier descriptors and Gaussian windowing for an interval of standard deviations (multiscale analysis), generating the curvegram. Figure 5 (a). and (b) exemplifies the high curvature points along the nucleus contour for a particular $\sigma$ and a threshold ($T=0.1$), which determines the minor value from which a curvature point is considered "high". The threshold parameter must be as low as possible in order to eliminate the noise inherent to spatial quantization effects. The threshold estimation occurred through successive experiments and user visualization of the relevant concavities of the addressed images.

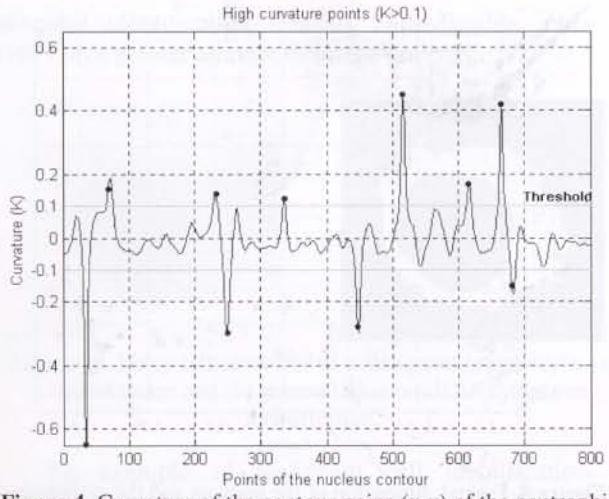


**Figure 4.** Curvature of the contour poins (x,y) of the neutrophil nucleus (Figure 5); the curve is smoothed for gaussian with $\sigma$=22.5. Black dots indicates the high curvature peaks of the contour.

An important descriptor originates from the collection of curvatures in terms of the scales (curvegram) – the bending energy [15], expressing the amount of energy needed to transform the specific shape under analysis into its lowest energy state (a circle). The bending energy curve represents the evolution of the energy along a $\sigma$ interval, i.e., a high curvature analysis of points in a multiscale approach. To achieve scale independence, the bending energy must be normalized by the perimeter ($L$) since we intend a scale-invariant shape analysis, which can be done by calculating the equation (4). Figure 5.c. presents the normalized bending energy ($B$) of the n points of the contour to an interval of $\sigma$.

$$B = L^2 / n \sum_{i=1}^{n} k(t,\sigma)^2 \qquad (4)$$

Another adopted concept is the image skeleton (Figure 6), which was calculated through a simple algorithm for exact dilations that propagate labels assigned to each contour point [6]. The multiresolution skeletons are obtained from the propagated labels and its higher values compose a skeleton [8]. The skeletonization relates to the minimal structure of the image, informing hierarchies, branches and angles among branches of this image. It can also be used for describing roughness, elongation of the object, high curvature points through counting the branches, calculation of the skeleton length and the number of extremities, respectively.

Fractal dimension can be used to express the complexity of an object, presenting how much the object samples the space around it, what we can characterize in terms of the effective surface of contact between the object and its surroundings. The fractal dimension estimation relates the area of the object to its rate of occurrence. The algorithm to calculate the multiscale fractal dimension curve [7] using distance transform [6] involves the following steps:

1. dilate the contour through distance transform using radii ($d$) between 0 to 200, keeping the area of the dilated objects;
2. construct a histogram relating the area with its respective radii ($h(d)$);
3. calculate the logarithm of the areas in terms of the logarithm of the radii (loglog);
4. derive the loglog function, obtaining the curve $C$;
5. calculate 2-$C$, the multiscale fractal dimension (Figure 7).

It should be observed that the $n$ points of the loglog plot extremities should be disconsidered due to poor sampling in the beginning of the curve and the discontinuity at the end of the curve. The fractal dimension peak indicates the maximum fractality achieved by the object. The step 1 is optimized in our program, once we keep the distance transform of the image during the exact dilation calculus for the skeletons.

## Results

Sabino et al [14] described preliminary measurements, including the general features of this current research.

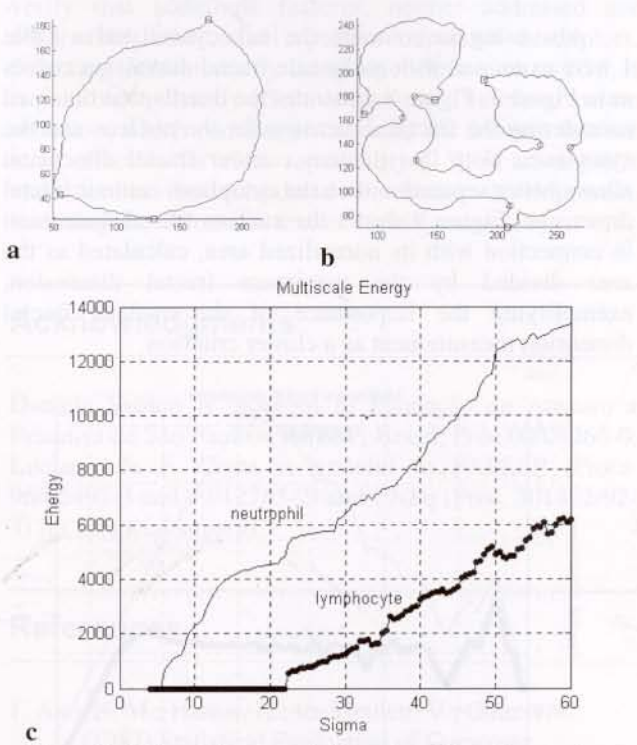


**Figure 5.** (a) Lymphocyte and (b) neutrophil nucleus contour with marked high curvature points for $\sigma$=22.5. (c) Bending energy curve of the lymphocyte (thick line) and the neutrophil (thin line) nucleus.

Original results of the application of methods as curvegram, multiscale skeletons and multiscale fractal dimensions to extract features are addressed.

Figure 5 (a) and (b) exhibit the contour of the nucleus of a lymphocyte and a neutrophil, respectively, and their high curvature points, assuming a specific threshold (high curvature = $k$>0.1) and $\sigma$=22.5. The number of concavities and convexities can be used to differentiate the lymphocyte from the neutrophil, for example. Figure 5 (c) shows the normalized bending energy, which allows inferring whether the lymphocyte contour is less bent than the neutrophil. The skeletons (Figure 6) of the two leukocyte contours, introduced in Figure 5, were calculated using the threshold $T$>15, which determines the skeleton ramification.

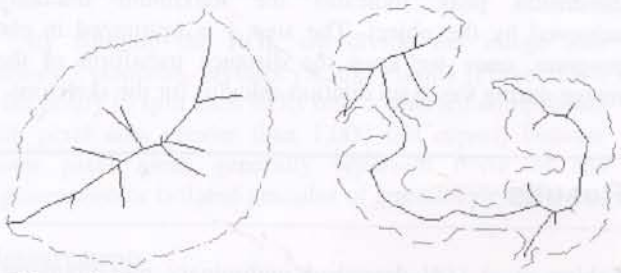

**Figure 6.** Skeletons (continuous line) of the lymphocyte (left) and neutrophil (right) nucleus contours (dashed line).

Also using the contours, the leukocytes listed in Table 1 were examined with multiscale fractal dimension curves as in Figure 7. Figure 8 illustrates the distribution obtained considering the fractal dimension for the nucleus and the cytoplasm. Note that nucleus contour fractal dimension allows better separation than the cytoplasm contour fractal dimension. Figure 9 shows the nucleus fractal dimension in connection with its normalized area, calculated as the area divided by the maximum fractal dimension, exemplifying the importance of the nucleus fractal dimension measurement as a cluster criterion.

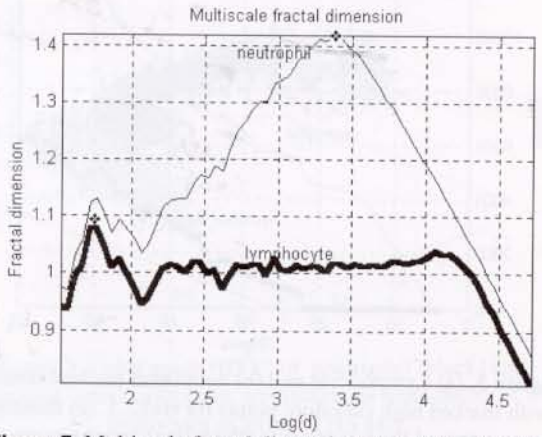


**Figure 7.** Multiscale fractal dimension curve of leukocytes contour (Figure 5): estimation of the maximal fractal dimension (black star).

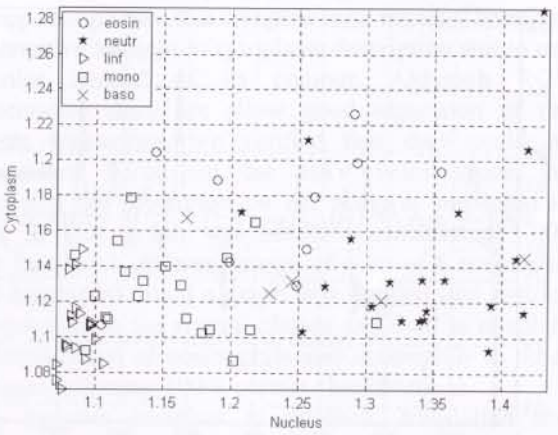


**Figure 8.** Fractal dimension of the nucleus and the cytoplasm: emergence of the division between *polimorphonucleateds* (circle=eosinophil, star=neutrophil and x=basophil) and *mononucleateds* (triangle=lymphocyte, square=monocyte) cells.

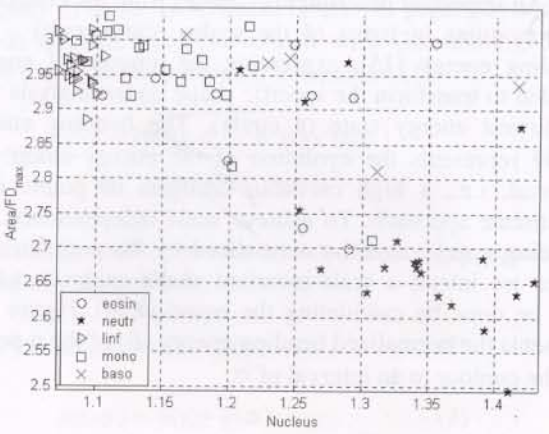


**Figure 9.** Fractal dimension of the nucleus in terms of its normalized area.

Although the nucleus fractal dimension allows visualization of a division between polimorphonucleateds and mononucleateds, the different types of leukocytes present a high scatter within their classes, mainly among polimorphonucleateds. The polimorphonucleated leukocytes contain lysosome granules in cytoplasm, often implying overlapping between the granules and the contour of nucleus and cytoplasm, the subject of our analysis.

A valuable feature to solve this problem would be the color of the cytoplasm, since specialists argue about the validity of the eosinophil color as a characteristic of its cytoplasm. Therefore, a cytoplasm color map was elaborated for neutrophils and eosinophils (Figure 3) and we verified that they could be differentiated using the hue-saturation axis. These two types of leukocytes were selected because their contours are often similar. The

basophil identification remains unpredictable, once it relies on a greater amount of image samples.

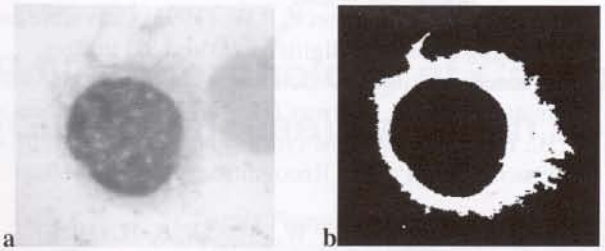

**Figure 10.** Hairy cell identified as a malignant lymphocyte. (a) original image and (b) segmentation result for cytoplasm identification.

An example of malignant cell identification is illustrated for the hairy cell in Figure 10 (a), whose cytoplasm is shown in Figure 10 (b). The identification begins with the measurements of roundness of nucleus, which is characterized by its circularity and low curvature points along the contour. So the algorithm recognizes the cell as a possible lymphocyte. An analysis of the cytoplasm in terms of its fractal dimension will indicate it is equal to 1.1714, which is higher than for a normal lymphocyte (fractal dimension varying between 1.076 to 1.149) as shown in Figure 8.

## Discussion

While research on leukocyte differentiation has been concentrated on graylevel images and thresholding [9,16], a method using color images has been considered here. Features such as curvature, skeletons and multiscale fractal dimension are promising for describing leukocytes. The analysis of the multiscale fractal dimension as a differentiation criterion was relevant to differentiate mononucleated from polimorphonucleated leukocytes, although segmentation improvements can lead to better results. The extracted features were first treated by using thresholding to classify the cells, though a clustering method [11] seems necessary. While only the nucleus curvature was analyzed, improvements in the segmentation process will allow including the curvature of the cytoplasm, which can help in blast cells identification, particularly hairy cells and cytoplasm membrane fragility. Some results [1,3,9] indicated that the contour and texture represent valuable parameters in the automatic leukemia diagnosis using cell morphology.

## Conclusion

Nowadays blood cell morphology has been progressively replaced by new and expensive technologies. However many hematologists can still diagnose some classes of leukemia just looking at the cell morphology, the cheapest way of analysis.

The current paper presented a collection of different measures applied to images of white blood cell nucleus and cytoplasm for evaluating its description in terms of the extracted features. Some contributions related here were the HS color space mapping of neutrophil and eosinophil cytoplasm, multiscale analysis of leukocyte contour considering curvature, skeletons and fractal dimension as well as the normalized bending energy. The described features showed encouraging results in the leukocyte differentiation, reporting a possible procedure for Hairy cell recognition, an abnormal and malignant type of cell.

The training set for segmentation and Gaussian probability distribution function of the RGB image pixels determine the precision of the results, observing worse classifications using the training set from different types of leukocytes. More samples should be considered and improvements to circumvent the problem of multimodal data representation of the color points should include estimation of probability density functions using Parzen window [11].

The feature extraction supplies the measurements of the cell, but few of them, usually, may be relevant descriptors to differentiate cells. It is also possible to verify that additional features, neither addressed nor observable before, can represent important descriptors. Selecting the features by using algorithms such as FSS, it is possible to identify effective features. Therefore, feature selection, which consists of choosing the features that are most effective for class separability [13], is an essential endeavor to be considered in future developments, including the chromatin texture investigation [16].

## Acknowledgments

Daniela Sabino is indebted to Fundação de Amparo a Pesquisa de São Paulo, FAPESP, Brasil, Proc.00/08266-0. Luciano da F. Costa is grateful to FAPESP (Procs. 96/05497-3 and 99/12765-2) and CNPq (Proc. 301422/92-3) for financial support.

## References

1. Aus, H. M.; Harms, H.; ter Meulen, V.; Gunzer, U. (1987) Statistical Evaluation of Computer Extracted Blood Cell Features For Screening Population to Detect Leukemias (ed. Devijver, P.A. and Kittler, J.) Pattern Recognition Theory and Applications, F 30:509-518.

2. Cesar, Jr.R. M. & Costa, L. da F. (1997) Application and assessment of multiscale bending energy for morphometric characterization of neural cells, Rev Sci Instrum, 68(5):2177-2186.

3. Cesar, Jr.R. M.,& Costa, L. da F. (1995) Piecewise linear segmentation of digital contours in O(N.LOG(N)) through a technique based on effective digital curvature estimation, Real-Time Imaging, 1:409-417.

4. Comaniciu, D.; Meer, P.; Foran, D.J. (1999) Image-guided decision support system for pathology. Mach Vision Appl, 11:213-224.

5. Costa, L. da F. & Cesar, Jr. R.M. (1996) Towards effective planar shape representation with multiscale digital curvature analysis based on signal processing techniques. Pattern Recogn, 29(9):1559-1569.

6. Costa, L. da F. & Cesar, Jr. R.M. (2001) Shape Analysis and Classification, Theory and Practice, CRC Press.

7. Costa, L. da F.; Campos, A.G.; Manoel, E. T. (2001) An integrated approach to shape analysis: results and perspectives, International Conference on Quality Control by Artificial Vision, Le Cresout, France.

8. Costa, L. da F., Estrozi, L. F.(1999) Multiresolution Shape Representation without Border Shifting. Electronics Letters 35(21):1829-1830.

9. Dias A. V. (1995) Avaliação de classificadores Bayesianos para identificação de células cancerígenas em imagens microscópicas, Ms. Thesis.

10. d'Onofrio, G., Zini, G., Bain, B. (1996) Morphology of the blood. Butterworth Heinemann.

11. Duda, R., Hart, P.E., Stork, D. (1997) Pattern Classification, 2nd edition.

12. Estrozi, L.F., Campos, A.G., Rios, L.G., Cesar, Jr. R.M., Costa, L. da F. (1999) Comparing curvature estimations techniques, 4° SBAI – Simpósio Brasileiro de Automação Inteligente, SP.

13. Fukunaga K. (1990) Introduction to Statistical Pattern Recognition, Academic Press, 2nd edition.

14. Sabino, D.M.U., Costa, L.da F., Martins, S.L.R., Zago, M.A. (2001) Diferenciação de leucócitos por computador, XVIII Congresso Nacional do Colégio Brasileiro de Hematologia.

15. van Vliet, L. J. & Verbeeck, P.W. (1993) Curvature an Bending Energy in digitized 2D and 3D images, Proceedings of the 8th Scandinavian Conference on Image Analysis, Norway (Hogda, K. A., Braathen, B., Heia, K.), NONIM-Norwegian Soc. Image Process. and Pattern Recognition, 2:1403-1410.

16. Young, I.T.; Verbeeck, P.W.; Mayall, B. H. (1986) Characterization of chromatin distribution in cell nuclei, Cytometry 7:467-474.

17. Young, I.T. & Roos, R. (1988) Acuity: image analysis fo the personal computer, Pattern Recognition and Artificial Intelligence (ed. Gelsema E.S. and Kanal L. N.) Elsevier Science Publishers, 7:5-16.