Clustering as a Method of Image Simplification

Anna Korzynska and Mateusz Zdunczuk

Laboratory of Microscopic Image Processing Information Systems, Department of Hybrid Biosystems Engineering, Polish Academy of Sciences Institute of Biocybernetics and Biomedical Engineering, 4 Trojdena Str., 02-109 Warsaw, Poland
akorzynska@ibib.waw.pl

Summary. The microscopic images of the cells are very difficult to analyze and to segment. The advanced method of segmentation such as region growing, watershed or snake requires the initialization information about the rough position of the cell body. It is proposed to localize cells in image using a threshold of simplified image. Clustering grey levels in image is proposed to simplify image. The k-means clustering method supported by weighting coefficients is chosen to collect all grey tones presented in the background into one cluster and other grey tones into few clusters in such a way that they cover a cell region in microscopic images. The weighting coefficients are used to influence (expand or contract) patterns in microscopic images of living cells. The method was evaluated on the basis of confocal and bright field microscopy images of cells in culture.

1 Introduction

The microscopic images of the cells are very difficult to analyze because of lack of precise and accurate methods of cells separation from the background. The segmentation of cell images are not easy due to the contrast quality, variation in cell shape, temporal changes in image contrast and focus, what is shown in Fig. 1.

The advanced method of segmentation such as region growing, watershed or snake requires the initialization information about the rough position of the cell body. The main idea of this research is to localize cells in image using one of the clustering methods. The k-means clustering method supported by weighting coefficients is proposed to reduce quantity of grey tones in image in such a way that the background variation is suppressed into one cluster and the other clusters cover cell area. This type of simplified image, after being thresholded, allows to localize the cell body fragments. Next, using mathematical morphology, binary operations, the cell fragments would be connected and holes in their area would be filled.

2 Related Research Review

Clustering has a rich history in pattern recognition [29], image processing [6, 10] and information retrieval [15, 16]. In this paper this methodology is employed in image processing as a low-level procedure that aims at simplifying an image.

E. Pietka and J. Kawa (Eds.): Information Tech. in Biomedicine, ASC 47, pp. 345–356, 2008.
springerlink.com

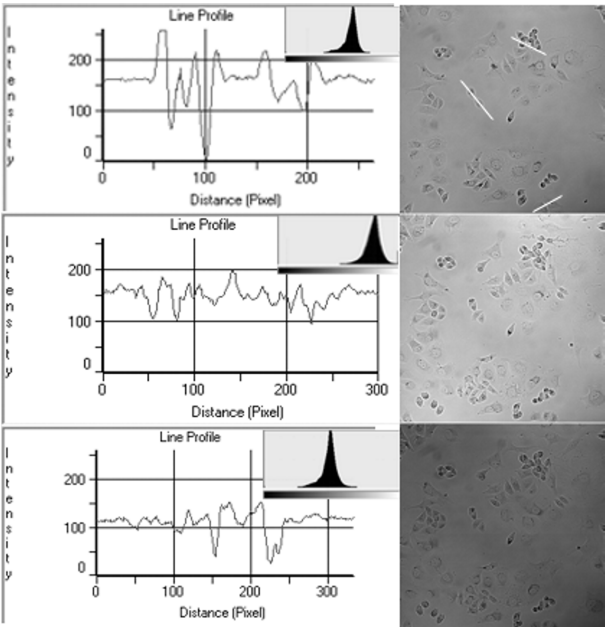

Fig. 1. Three images with their histograms present external source light variations and internal nonhomogenity in the light distribution. There are three line profiles: the top line profile shows intensity function along the top white line crossing the relatively well contrasted and small cell cluster, the central one shows intensity function along the middle white line crossing the middle size cells, and the last one shows intensity function a long bottom white line crossing large, flattened and poorly contrasted cell.

Analogical, but not the same aims have been investigated by Du et al. [8] and by Duda and Hart [9]. Both groups of researchers were concentrating on partitioning an image into homogeneous regions in the sense of segmentation rather then object localization. Among methods of image segmentation, using clustering, the most interesting are: k-means [13, 25], isodata [3], and fuzzy c-means [28]. In this investigation the variation of k-means method was chosen as very easy to adjust to particular image by manipulation of weighting coefficients.

The k-means clustering methods were introduced in 1967 by J. MacQueen as an unsupervised classification technique. These methods were used to detect cell nuclei in digital image-based cytometry [20, 26], but only for fluorescent microscopic images which are easier to analyze rather then bright field or confocal microscopy images. Since bright field and confocal microscopic images segmentation using such methods as watershed [21, 17], region growing [18, 24], model-based [22] and agent based or hybrid method [2, 4], gives good accuracy and precision but all these methods need initial information of the cell position, so the k-means clustering is proposed to be pre-processing phase of these images segmentation methods.

3 Microscopic Image Characterization

The microscopic images of cells are very difficult to analyze. This is because of the image quality which depends on a type of microscope, a type of cells and a resolution of acquired image [19, 7].

The bright field microscopy and the scanning confocal microscopy produce greyscale images with poorly contracted cells in the image plane, see Fig. 1. Cells are transparent objects so they transmit light and they are visible as grey tones which are darker or brighter then grey background. Some path of the cell body is in the background grey level range. Some cells are partly or fully rounded by halo which appears as brightened background. This brightness is caused by light reflection on cell wall. There is no halo and contrast between a cell and the background in the parts where cell is very flattened. Furthermore the intensity of the background is not uniform across the image, due to the external and source light variation.

Three graphs of line profiles in Fig. 1 present significant intensity changes across the image plane (bottom-right and top-left image corners are darker then bottom-left and top-right) in both, in the background and within the cell. The variations of the intensity caused by noise is also observed in the background and in the cell area. The histogram of each microscopic cell image is unimodal and slightly skew. The presented histograms are located in various positions of grey scale according to mean lighting conditions, see Fig. 1.

There is a difference in detail visibility according to microscopic techniques. Neural stem cells sample observed in the red laser confocal scanning microscopy and the bright field microscopy image are presented in Fig. 2. The bright field microscopy builds images (see right part of Fig. 2) with relatively large deep of field in comparison to the laser confocal microscopy (see left part of Fig. 2).

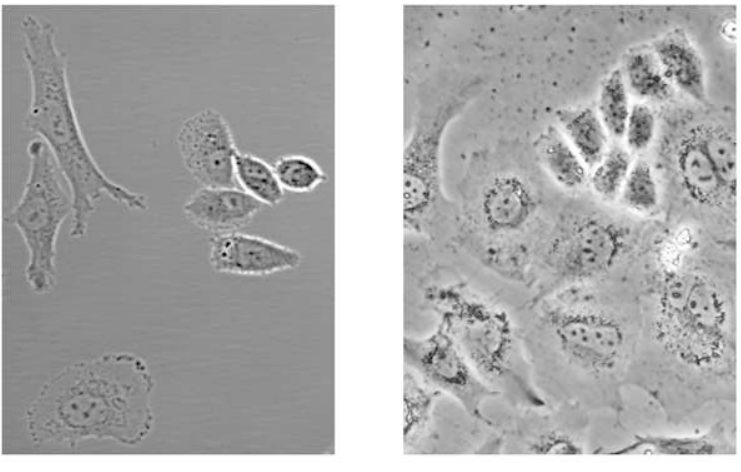

Fig. 2. Cells images in the red light laser scanning confocal microscopy (left) and the bright field microscopy (right)

The structures visible in confocal microscopy such as nucleus, the endoplasmic reticulum around nuclei only sometimes are observed in bright field images as blurred objects.

4 Methods

The goal of the study is to find the method of the microscopic image simplification which suppress the background variation into one grey tone cluster and cell regions in the other clusters. It is proposed to use k-means clustering method to do that.

4.1 Clustering Method

Clustering is a commonly used technique for features determination and extraction from large data sets and for the determination of similarities and dissimilarities between elements in the data sets.

In this paper the concept of image clustering is exploited. Image pixels are grouped in such a way that the information which image contains is emphasized or extracted. Generally, the criteria used to collect pixels in clusters are various, among them color, texture, connectivity, gradient and so on and they depend on image characteristics and objects in the image and on the goal of the image processing. Pixels collected in one cluster are presented by similar or the same color or grey level in simplified image what leads to image segmentation. This process reduces image noise and some image details. If these details carrying information which is redundant or not important according to the goal of the image processing, the image simplification does not damage image, but rather highlights its sense.

Mathematical Background

The clustering problem can be formally defined as follows [12, 27].

Given a data set $U = \{u_1, u_2, ..., u_p, ..., u_{N_p}\}$ where u_p is a pattern in the N_d-dimensional feature space, and N_p is the number of patterns in U, then the clustering of U is the partitioning of U into K clusters $\{V_1, V_2, ..., V_K\}$ satisfying the following conditions:

- Each pattern should be assigned to a cluster, i.e. $\bigcup_{k=1}^{K} V_k = U$
- Each cluster has at least one pattern assigned to it $V_k \neq \emptyset$ for $k = 1, ..., K$
- Each pattern is assigned to one and only one cluster $V_l \cap V_k = \emptyset$ for $l \neq k$.

Clustering can be defined with reference to an image [8, 23], than given an image u, let $U = \{u(i,j)\}_{(i,j)\in D}$, where $D = \{(i,j) : i = 1, ..., I, j = 1, ..., J\}$

for positive integers I and J, (i,j) are integer pairs that range over the image domain, denote the set of grey tones values in the original image. Then, for any set of the replacement grey tones $W = \{w_k\}_{k=1}^{K}$, called generators, let

$$V_l = \{u(i,j) \in U : |u(i,j) - w_l| \leq |u(i,j) - w_k|, k = 1, ..., K\} \tag{1}$$

$V_l, l = 1, ..., K$ denotes the subset of those values of the grey tones that are closest to w_l (in the sense of the euclidian distance in 1-dimensional space of grey levels) than to any of the other w_k's. The subset of grey tones V_l is called the cluster corresponding to w_l and the set of subsets $V = \{V_k\}_{k=1}^{K}$ is called a clustering of the set U of grey tones. For any non-overlapping covering of $U = \{V_k\}_{k=1}^{K}$ into K subsets, one can define the means or centroids of each subset V_k as the grey value $\bar{w}_k \in V_k$ that minimizes following expression called clustering energy

$$\sum_{k=1}^{K} \sum_{u(i,j) \in V_k} |u(i,j) - w_k|^2 \tag{2}$$

The grey values w_k that generate the clustering such that $w_k = \bar{w}_k$ for $k = 1, ..., K$ are called the centroids of the associated clusters.

Several variants of the k-means algorithm have been reported in the literature [1]. Some of them attempt to select a good initial partition so that the algorithm is more likely to find the global minimum value. Another variation is to permit splitting and merging of the resulting clusters. Typically, a cluster is split when its variance is above a pre-specified threshold, and two clusters are merged when the distance between their centroids is below another pre-specified threshold. Using this variant, it is possible to obtain the optimal partition starting from any arbitrary initial partition, provided proper threshold values are specified. Another variation of the k-means algorithm involves selecting a different criterion function. The weighted k-means algorithm is a variation of the classic k-means algorithm [14, 29]. Weight coefficients, which provide weighted distortions between data and cluster centers, are incorporated into the algorithm to realize anticipated clustering. One can redefine the energy expression Eq. 2 so that the contributions from each of the clusters are weighted. This allows, for example, for a given grey tone to be included in a large cluster and opposite. Applied to a digital image, weighted clustering can let grey tone generators focus on selected details of the image and not be overwhelmed by other grey tones. Definition of the weighted energy expression is as follows

$$\sum_{k=1}^{K} \lambda_k \sum_{u(i,j) \in V_k} |u(i,j) - w_k|^2 \tag{3}$$

where λ_k are positive weighting factors. In general, λ_k is allowed to depend on factors such as the cardinality $|V_k|$ of the subset V_k, the within cluster variance, etc. [14]. In the resulting image original grey levels are replaced by centroids of the last iteration of the proposed algorithm.

Algorithm

The k-means method aims to minimize the sum of squared distances (in the sense of grey levels) between all points and the cluster center. Algorithm works recursively:

1. Initialization phase:
   a) Choose K initial cluster centers $w_1, w_2, ..., w_K$.
   b) At the k-th iterative step, distribute the samples $u(i,j)$ among the K clusters using the relation,

$$u \in V_k \text{ if } |u(i,j) - w_l| \leq |u(i,j) - w_k| \text{ for } k = 1, ..., K \tag{4}$$

   where V_k denotes the set of samples whose cluster center is w_k and $l \neq k$.
2. Iterative phase
   a) Compute the new cluster centers $w_k(k+1)$, $k = 1, ..., K$ such that the sum of the squared distances from all points in V_k to the new cluster center is minimized. The measure which minimizes this is the sample mean value of V_k. Therefore, the new cluster center is given by

$$w_k(k+1) = \frac{1}{N_k} \sum_{u \in V_k} u \text{ for } k = 1, ..., K \tag{5}$$

   where N_k is the number of samples in V_k.
   b) If $w_k(k+1) \cong w_k(k)$ for $k = 1, ..., K$ (with respect to a chosen threshold value) then the algorithm has converged and the procedure is terminated. Otherwise go to step 2.

Different stopping criteria can be used in an iterative clustering algorithm:

- the change in centroid positions are smaller than a user-specified value,
- the quantization error is small enough,
- a maximum number of iterations has been exceeded.

In the proposed method the last stopping criterion is used and merging procedure supported by weighting coefficients are exploited. So the resulting cluster consists of clusters corresponding for grey levels of the background. The weighting coefficients are selected on the basis of the cluster size counted from the previous iteration.

4.2 Details of the Proposed Method

A major problem with k-means algorithm is its sensitivity to the selection of the initial generator number and their positions and its convergence to a local minimum of the criterion function if the initial partition is not properly chosen. In this investigation the number of the generators ranges from 5 to 25 and three various methods of choosing the initial generators were tested:

1. random choice,
2. homogenous choice,
3. arbitrary choice,

of the grey levels over grey scale. Resulting images for various values of K and for various methods of generators initialization are presented in Sect. 6.

5 Material

Evaluation of the proposed method was done using microscopic images of neural stem cell culture [5]. The images were acquired from digital cameras attached to two microscopes: bright field inverted microscopy (Olympus IX70 the right side Fig. 2) and scanning confocal microscope with the red color laser (Zeiss Fig. 1 and the left side Fig. 2). The observation plane on culture dishes cover 100 x 100 μm space in the first case while 120 x 120 μm in the second. Bright field microscopy images were acquired as a digital image of 1024 x 1024 pixels in 12 bits deep acquisition and next converted to 8-bit deep images by the linear resampling of the grey scale from minimum to maximum of grey levels. In the case of confocal microscopy images 8 bits deep acquisition of 2048 x 2048 pixels were done. Bicubic resampling was used to resample images to the size 1024 x 1024.

6 Results

The proposed method results on microscopic images were analyzed and compared in both qualitative and quantitative manner.

6.1 How Initialization Influences the Results

Fig. 3 shows the results of the proposed method with increasing number of clusters and with various generators distribution over grey scale.

It can be observed that the number of clusters influences the resulting image in the detail level and in the smoothing of the background. The larger cluster number, the more grey values are bounded up with the background. Therefore smoothing of the background is desirable, the choice of 8 clusters seems to be the best to achieve the smoothing of the background. The dependence of the results on the way generators are chosen isn't unambiguous. Because tested strategies of generators choice do not lead up to fundamental differences in the final distribution of the centroids after many iterations and difference among results are not ambiguous and not significant, arbitrary choice of generators positions was used in further investigation.

6.2 Evaluation of Influence of Selected Weighting Coefficients

The procedure that involves weighting coefficients aims at such partitioning of the picture grey levels that some pixels which grey tones correspond to the

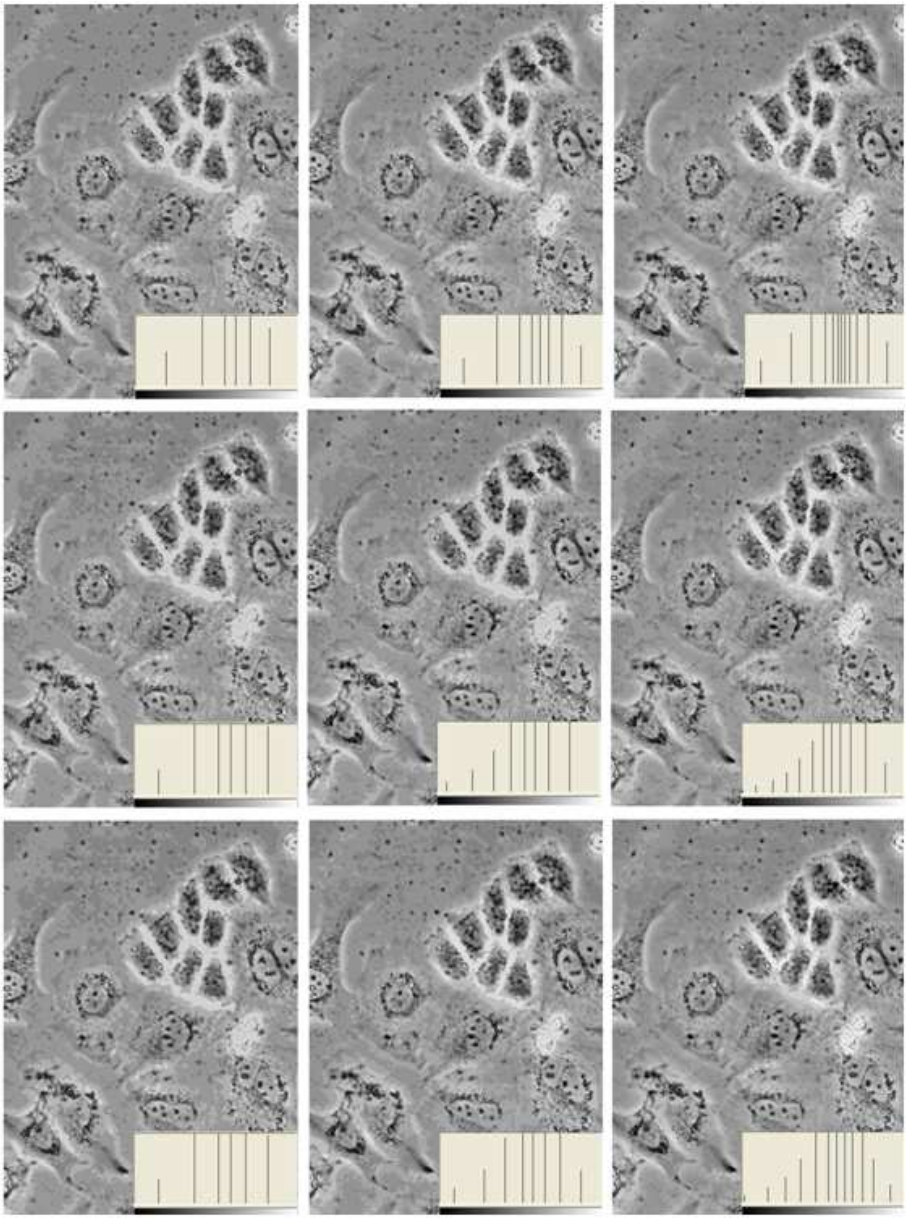

Fig. 3. Proposed clustering method results with corresponding image histograms for various initialization methods: with increasing number of clusters from $k = 6$ in the first column, by $k = 8$ in the second one to $k = 12$ for the last one and with starting centroids chosen as: randomly distributed across the grey scale in the first row, homogenously distributed in the second one and arbitrary chosen by operator in the third one

Table 1. Matching matrix resulting from weighted k-means algorithm: distribution of classified pixels percentages compared with classification without weighting coefficients

Image	Weighting coefficients	Centroids of the last iteration; Resulting centroids	Fraction of pixels displaced
1-row,1-column	0 0 0 0 0 0 0 0	32 73 98 108 119 140 177 239	—
2-row,1-column	150 0.01 0.01 2048 10456 5586 0.01 45	43 74 85 113 141 185	81,2%
2-row,2-column	150 654 856 0.01 0.01 0.01 75 130	43 74 85 95 113 141 245	45,7%
2-row,3-column	150 654 856 2048 10456 5586 75 130	43 74 85 95 113 141 185	62,1%

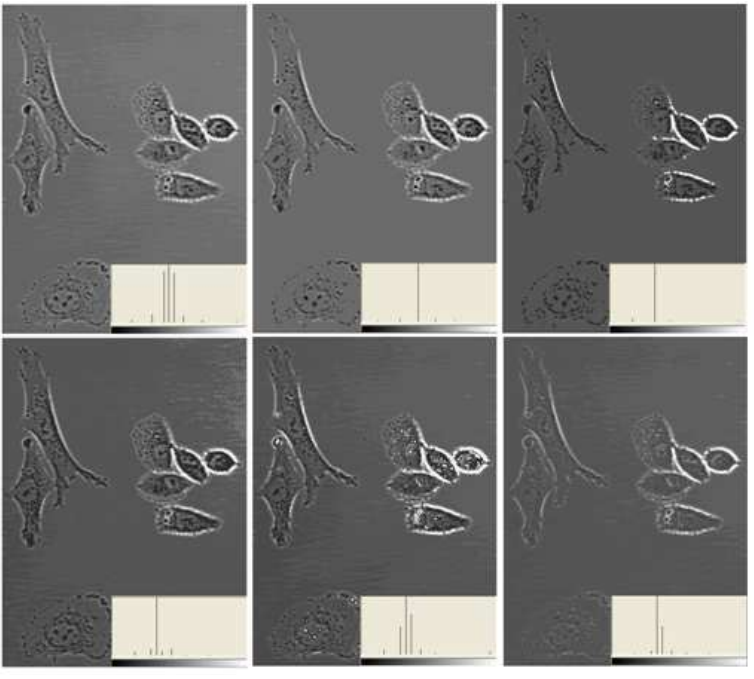

Fig. 4. The results of the proposed clustering method for arbitrary chosen generators position as described in the text with corresponding image histograms: - in the first row (from left to right): reference image with all weighting coefficients equally influences all clusters (equal 1); reference image with merged background grey levels (98 and 119 merged to 108); image with weighting coefficients which influence five first clusters [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 1, 1]; in the second row images with coefficients calculated based on cardinality of the chosen clusters (from left to right): resulting in the homogeneity of the background, resulting in the expansion of the objects, resulting in both homogeneity of the objects and homogeneity and the expansion of the background

one class are placed in the other class. Generally, pixels are attracted by other classes if coefficients of these classes are small numbers. This mechanism is used

to include background pixels around cells structures to the cell classes to achieve a cohesive object region. Matching matrix, presented in Tab. 1, was calculated to evaluate fraction of pixels reclassified from one class to the other one according to a chosen set of coefficients. Classification without weighting coefficients, which is presented in both the first row in the Tab. 1 and in Fig. 4 the first row and the first column image, is used as a reference to calculate pixels displacements. The last three rows of Tab. 1 present the matching matrix results for images shown in the second row in Fig. 4.

6.3 How Cardinal Coefficients Influences the Results

Fig. 4 shows the results of the proposed method with 8 clusters and arbitrary chosen positions of initialization generators: [42 68 86 109 137 149 162 175] and with various values of weighting coefficients. In this test weighting coefficients different than equal to 1 appeared for various clusters to investigate influence of specific cluster expansion or contraction in the resulting image.

It was observed that in order to narrow the area of the determined cluster, the coefficients are selected to be equal to the cluster size. Otherwise, in case of the cluster area expansion, coefficients are selected as the inverse of the cluster size. The results show that choosing such coefficients that are leading to the background expansion (first image in the second row), gives the best effects while the grey tones present in cell area are manipulated the results images are worse (middle and last images in the second row). The best result was obtained by merging procedure (middle image in the first row).

7 Discussion and Conclusion

The proposed greyscale image simplification method described in this paper employs clustering method to redefine grey value of each pixel in microscopic images in such a way that the background pixels are collected into one cluster and the other clusters collect pixels of other grey tones. Resulting images, presented in the text, show that the used method is less dependent on the way the starting centroids are chosen rather than on the number of these centroids and that weighting coefficients based on the cluster cardinality allow manipulation of the cluster size. These studies conclude that the proposed method is promising enough to carry out the influence on the use of another types of coefficients. It seems that texture would be a good choice for a new coefficient definition because texture is the feature which discriminate area of the cell from the background area.

Acknowledgement. We are grateful for cell impart from line HUCB-NS to the experiments and for the support in experiments received from the NeuroRepair Department Laboratory, Polish Academy of Sciences Medical Research Center.

References

1. Anderberg, M.: Cluster Analysis for Applications. Academic Press, New York (1973)
2. Baujard, O., Garbay, C.: KISS: a multiagent segmentation system. Optical Engineering 32(6), 1235–1249 (1993)
3. Bezdek, J.C.: A Convergence Theorem for The Fuzzy ISODATA Clustering Algorithms. IEEE Transaction On Pattern Analysis And Machine Intelligence 2(1), 1–8 (1980)
4. Boucher, A., Doisy, A., Ronot, X., Garbay, C.: Cell Migration Analysis Afte. Vitro Wounding Injury with a Multi Agent Approach. Artificial Intelligence Review 12, 137–162 (1998)
5. Buzanska, L., Jurga, M., Stachowiak, E.K., Stachowiak, M.K., Domanska–Janik, K.: Focus on Neural Stem Cells. Neural Stem-Like Cell Line Derived from a Nonhematopoietic Population of Humane Ubilical Cord Blood. Stem Cell and Development 15, 391–406 (2006)
6. Castleman, K.: Digital Image Processing. Prentice Hall, Englewood Cliffs (1996)
7. Comaniciu, D., Meer, P.: Cell image segmentation for diagnostic pathology. In: Suri, J.S., Setarehdan, S.K., Singh, S. (eds.) Advanced algorithmic approaches to medical image segmentation: state-of-the-art application in cardiology, neurology, mammography and pathology, pp. 541–558 (2001)
8. Du, Q., Faber, V., Gunzburger, M.: Centroidal Voronoi tessellations: Applications and algorithms. SIAM Rev 41, 637–676 (1999)
9. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley and Sons, New-York (1973)
10. El-Sakka Mahmoud, R., Kamel Mohamed, S.: Adaptive Image Compression Based on Segmentation and Block Classifcation. Int. Journal of Imaging Systems and Technology 10(1), 33–46 (1999)
11. Garbay, C., Chassery, J.M., Brugal, G.: An interactive region growing process for cell image segmentation based on local color similarity and global shape criteria. Anal. Quantit. Cytol. Histol. 8, 25–34 (1986)
12. Hartigan, J.: Clustering Algorithms. Wiley Interscience, New York (1975)
13. Hartigan, J., Wong, M.: Algorithm AS 136: A k-means clustering algorithm. Appl. Stat. 28, 100–108 (1979)
14. Inaba, M., Katoh, N., Imai, H.: Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In: Proc. Tenth Ann. Symp. on Computational Geometry, pp. 332–339 (1994)
15. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
16. Jain, A., Murty, M., Flynn, P.: Data Clustering: A Review. ACM Computing Surveys 31(3), 264–323 (1999)
17. Jiang, K., Liao, Q.M., Dai, S.Y.: A novel white blood cell segmentation scheme using scale-space ltering and watershed clustering. In: Proc. Int. Conf. on Machine Learning and Cybernetics, vol. 5, pp. 2820–2825 (2003)
18. Liao, Q., Deng, Y.: An accurate segmentation method for white blood cell images. In: Proc. Int. Symposium on Biomedical Imaging, pp. 245–248 (2002)
19. Liedtke, C.E., Gahm, T., Kappei, F., Aeikens, B.: Segmentation of microscopic cell scenes. Analyt. Quant. Cytol. Histol. 9, 197–211 (1987)
20. Lockett, S.J., Herman, B.: Automatic detection of clustered, fluorescent-stained nuclei by digital image-based cytometry. Cytometry 17, 1–12 (1994)

21. Malpica, N., Ortiz, C., Vaquero, J.J., Santos, A., Vallcorba, I., García-Sagredo, J.M., Pozo, F.: Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry 28, 289–297 (1997)
22. Nilsson, B., Heyden, A.: Model-based segmentation of leukocyte clusters. Proc. Int. Conf. on Pattern Recognition 1, 727–730 (2002)
23. Okabe, A., Boots, B., Sugihara, K.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. Wiley, Chichester (1992)
24. Ongun, G., Halici, U., Leblebicioglu, K., Atalay, V., Beksac, M., Beksak, S.: An automated differential blood count system. In: Proc. Int. Conf. of the IEEE Engineering in Medicine and Biology Society, vol. 3, pp. 2583–2586 (2001)
25. Phillips, S.J.: Acceleration of k-means and Related Clustering Algorithms. LNCS. Springer, Heidelberg (2002)
26. Proffitt, R.T., Tran, J.V., Reynolds, C.P.: A fluorescence digital image microscopy system for quantifying relative cell numbers in tissue culture plates. Cytometry 24, 204–213 (1996)
27. Rasmussen, E.: Clustering Algorithms. In: Frakes, W.B., Baeza-Yates, R. (eds.) Information Retrieval: Data Structures and Algorithms, Prentice Hall, Englewood Cliffs (1992)
28. Zadeh, L.A.: Fuzzy Sets. Inform. Control 8, 338–353 (1965)
29. Zhang, Y.J.: A Survey on Evaluating Methods for Image Segmentation. Pattern Recognition 29(8), 1246–1335 (1996)