SVM-Based Cancer Grading from Histopathological Images Using Morphological and Topological Features of Glands and Nuclei

Catalin Stoean, Ruxandra Stoean, Adrian Sandita, Daniela Ciobanu, Cristian Mesina and Corina Lavinia Gruia

Abstract The paper puts forward a new data set comprising 357 histopathological image samples obtained from colon tissues and distinguished into four cancer grades. At the same time, it proposes an automatic methodology for extracting knowledge from these images and discriminating between the disease stages on its base. The approach identifies the glands and nuclei and uses morphological and topological features related to these components to generate 76 attributes that are further used for classification via support vector machines. The values of one parameter used for the identification of the nuclei are tuned and surprisingly good results are reached when overlapping nuclei are identified as singular objects.

Keywords Image processing · Histopathological image · Feature extraction · Classification

C. Stoean (✉) · R. Stoean · A. Sandita
Faculty of Sciences, Department of Computer Science,
University of Craiova, Craiova, Romania
e-mail: catalin.stoean@inf.ucv.ro; cstoean@inf.ucv.ro

R. Stoean
e-mail: ruxandra.stoean@inf.ucv.ro

A. Sandita
e-mail: asandita@inf.ucv.ro

D. Ciobanu · C. Mesina
Faculty of Medicine, University of Medicine and Pharmacy of Craiova,
Craiova, Romania
e-mail: elada192@yahoo.com

C. Mesina
e-mail: mesina.cristian@doctor.com

C.L. Gruia
Emergency County Hospital of Craiova, Craiova, Romania
e-mail: paraschivdan65@yahoo.com

G. De Pietro et al. (eds.), *Intelligent Interactive Multimedia Systems and Services 2016*, Smart Innovation, Systems and Technologies 55,
DOI 10.1007/978-3-319-39345-2_13

1 Introduction

Histopathological images are obtained by staining body tissues that are subsequently examined under a microscope [1]. The current protocol for cancer diagnosis in general, and for the colorectal type, in particular, involves the human expert (pathologist) analysis of the histopathological image while the best treatment for cancer consists in its early diagnosis [2, 3]. Although the judgement of the pathologist is educated and based on a vast experience, it is subjective and may lead to serious variability [1, 4, 5]. Furthermore, the physicians are confronted with a vast amount of histopathological images, due to the important investments in developing advanced microscopy hardware and the persuasion of individuals to have medical examinations more often. A potential aid could come from quantitative image-based evaluation of digital pathology slides, as these could at least eliminate the most obvious cases and leave the pathologists to concentrate on the most difficult records.

There is an acknowledged lack of benchmark data sets in histopathology imaging for validating techniques and for their comparison. In this respect, a data set of 357 histopathological images is made available [6] and a diagnosis support methodology is proposed, which achieves an overall prediction accuracy of almost 80 % on distinguishing between cancer stages from information automatically extracted from the collection. Next section briefly describes the image data set and the proposed methodology to extract the features and set a diagnosis, Sect. 3 outputs the results and Sect. 4 contains concluding remarks.

2 Feature Extraction and Classification

The data set [6] contains histopathological images of normal tissue and for cancer grades G1, G2, and G3. For ease of reference, we will refer to them in the current paper as G0, G1, G2 and G3. The next subsection contains a detailed description of the collection. Subsequently, the feature extraction stage is concentrated on the identification of glands and nuclei and several measures derived from this information are further used to construct a numerical data set that is fed to a support vector machine (SVM) in order to accurately grade the tissues.

2.1 Histopathological Image Data Set

The images are obtained from colon tissue slides stained using hematoxylin and eosin (H&E) and obtained from an electron microscope at the ×10 magnification level at the Emergency County Hospital of Craiova, Romania. We depart from 30 histopathological images that represent cancer grades G1, G2 and G3 from 30 different patients. Most of them contain border regions between normal tissues and

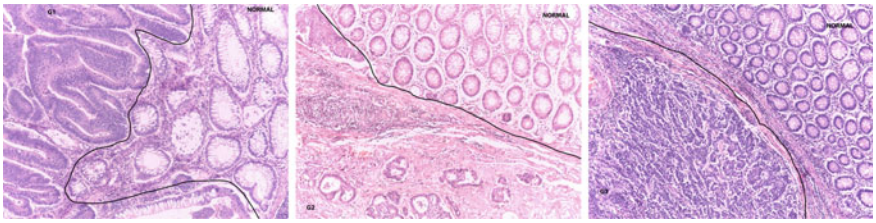

Fig. 1 Samples of the initial histopathological images. The borders are delineated between normal condition and grades 1, 2 and 3, respectively (from *left* to *right*)

malignant ones (see Fig. 1), so as in [7] representative parts could be extracted for G0, G1, G2 and G3. The obtained samples have a similar resolution of 800×600 pixels and they are 357 in total, distributed as follows: 62 cases as normal, 96 as G1, 99 as G2 and 100 as G3, respectively.

The images have various intensities, but they are all included in normal cases, as these are obtained from images contained in the representatives from the grades G1–G3.

2.2 *Feature Extraction and Proposed Diagnosis Support Methodology*

Usually, after the histopathological images are produced, computer-based grading follows several stages, i.e. image preprocessing, feature extraction, reduction of the number of feature and finally classification [1, 3–5, 8].

Although the images possess very different particularities, especially when the comparison is between distinct grades, the glands and the nuclei appear in most of them (there are exceptions with glands that are missing in some images denoting cancer grade 3). There are not many works that take into account both the glands and nuclei at the same time. Most of them concentrate on the nuclei and additionally tackle other textural features [3, 4, 7]. As opposed to the other studies, the focus of the current work thus becomes to detect and use these two common presences alike. For each found component, besides its counting, several characteristics were measured: the area and perimeter were computed, the enclosing circle was found and its radius was also taken into consideration. For every image and for the three measures the average, median, standard deviation and the ratio between the minimum and the maximum value were subsequently calculated.

Efforts for accurately identifying the gland interiors have been previously made ([9] and [10]) and the procedures that proved successful were employed in the current study. Gaussian smoothing is used as a preprocessing step as it showed to be more efficient as opposed to box, normalized box, median or bilateral filtering. It is then followed by a watershed algorithm [11] for boundary-based segmentation

of the glands interiors. The watershed algorithm uses a set of marked pixels on the grayscale version of the image for avoiding the creation of too many contours. For that, a thresholding operator is used to transform the image into a binary one. Then, noise elimination has to take place on the black and white image, i.e. the discarding of very small white spots on the black background (erosion), as well as the opposite (dilation). In [9] and [10] an evolutionary algorithm is used to search for good input parameter values for the described methodology and among them the threshold, number of erosions and dilations were included. In the current work, values for these parameters were chosen as the ones that previously proved to be well-suited overall.

Besides the statistical measures referring to the area, perimeter and radius of the glands, the layout of the components is also considered via two graph-based techniques. Consequently, the interior points of the enclosing circles for the detected glands (seeds) are used to draw the Delaunay triangles and Voronoi diagrams. The Delaunay triangles have the property that no point in the initial set lies inside the circles that enclose the formed triangles. The Voronoi diagrams conduct a separation of the plane into regions where all included points are closer to the specific seed of their regions as opposed to the other ones. Euclidean distance is used in the current implementation. An example of the application of the two techniques for four images containing grades G0–G4 is illustrated in Fig. 2. For the obtained triangles and polygons once again the area, perimeter and radius of the circumcircle are calculated and the same average, median, standard deviation, and minimum to maximum ratio are considered for each one of them.

In the current work, various methods had been tested for identifying the nuclei [3, 4, 7], like color quantization, k-means, fitting ellipses to the found contours via Hough Transform, the watershed algorithm and different preprocessing techniques had been tried (various filtering options, the Laplacian to sharpen the image). However, experiments showed that all these in multiple combinations conducted to sub-

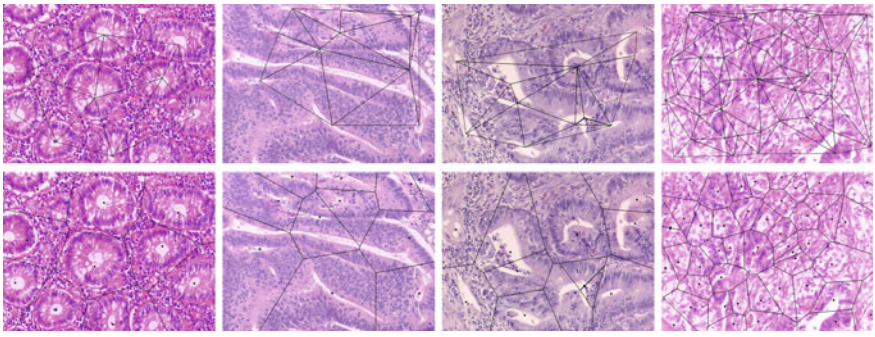

Fig. 2 Delaunay triangles (*first line*) and Voronoi diagrams (*second line*) found for the detected glands in images corresponding to a normal sample, grades 1, 2 and 3 (from *left* to *right*). The points that serve as inputs for the Delaunay triangles and Voronoi diagrams are found as central points for the *circles* that enclose the contours of the detected glands

optimal results as far as the visual appreciation of the physicians is concerned. Contours were obtained with higher precision when the distance transform algorithm [12] was applied for this purpose, as the method deals relatively well with overlapping nuclei [13–15].

The procedure flows as follows:

- The image is transformed into the grayscale variant.
- In a preprocessing step, the image is blurred using a normalized box filter.
- A threshold is applied to the obtained picture in order to obtain a binary one. As noticed in the experimental phase, the accuracy of the methodology depends at a high degree on the choice of this threshold parameter.
- The distance transform is applied on the resulting binary picture. The procedure calculates for every pixel the distance to the closest black one and thus produces a new image with the same size as the initial input. In our experiments an Euclidean distance is employed.
- The new image is then normalized.
- The image is later transformed into a binary one. On the resulting image, the contours of the detected objects represented the found nuclei.

Analogously to the case of gland detection, the area, perimeter, radius of the circumcircle of each detected object are computed and the associated statistical information associated to these are also calculated. The distribution of the found contours is also measured via the Delaunay triangulation and Voronoi diagrams with all the associated measurements as in the case of the glands. An example of the application of the two graph-based techniques for the detected nuclei can be seen for all the different grades in Fig. 3.

Table 1 recapitulates the features that are extracted from each image. As the number of attributes is relatively high, a SVM was employed to deal with the problem, as the methodology is acknowledged to be independent of the data dimensionality in

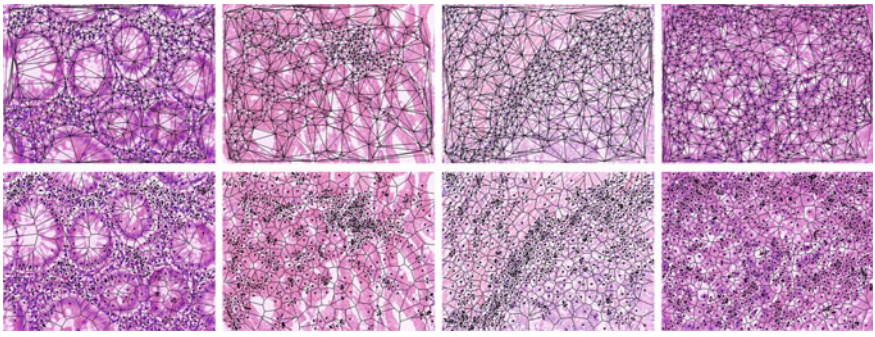

Fig. 3 Delaunay triangles (*first line*) and Voronoi diagrams (*second line*) found for the detected nuclei in another set of images corresponding to a normal sample, grades 1, 2 and 3 (from *left* to *right*). The points that serve as inputs for the Delaunay triangles and Voronoi diagrams are found as central points for the *circles* that enclose the contours of the detected nuclei

Table 1 The 76 features that are extracted from each histopathological image that enter into the subsequent classification process

Feature	Measures	Statistics	Total
Morphological	Area, perimeter, radius for glands and the same three for nuclei	Average, median, standard deviation, min/max	24
	Number of glands and of nuclei	–	2
Topological	Area, perimeter, radius for the Delaunay triangles and the same three for the Voronoi polygons for glands and repeated for nuclei	Average, median, standard deviation, min/max	48
	Number of Delaunay triangles for glands and for nuclei	–	2

The statistics are calculated for each measure on the line, for both glands and nuclei

a decision problem [16]. However, Principal Component Analysis (PCA) for reducing the number of features was also applied and SVM is afterwards utilized on the resulting data.

3 Experimental Results

The considered classification problem contains four classes to distinguish between, is characterized by 76 numerical attributes and contains 357 samples. Each sample is obtained from one histopathological image and the numerical values depend very much on the parameter values set for the procedures that derived them. In the current experiments, the interest lies in achieving a good overall accuracy, finding the accuracy between each pair of grades taken in turn, but also in observing how much does one parameter count (i.e. the threshold applied prior to the distance transform) in the overall diagnosis process.

3.1 Task

Observe the effect of the threshold parameter used to identify the nuclei over the overall accuracy of the classifier.

3.2 Setup

In the pre-experimental tests, it was observed that the SVM with a linear kernel conducted to significantly better results than when using a radial one. The threshold

was set to 160, as the physicians visually appreciated the results on several images as reasonable. The PCA reduced the number of attributes from 76 to only 13, as this was the number of components required to explain at least 95 % of the variance. For the PCA-transformed data, the SVM with a radial kernel reached an accuracy that was around 4 percent better than when using a linear kernel.

The threshold that transforms the image into a binary one prior to the application of the distance transform value is tuned from 130 to 200, considering only multiples of 5. Therefore, there are 15 different numerical data sets for the histopathological images and each one is subject to classification by SVM.

The numerical data set is randomly split into 2/3 training samples that are used to instruct the SVM classifier and 1/3 test data that is used for computing the accuracy. The process is repeated 30 times in order to verify the significance of the results.

The overall accuracy is computed as the percent of samples in the test set that are correctly labeled. For a deeper investigation of the results, each grade is considered in turn as opposed to the other ones and several insights are made available through accuracy, sensitivity, specificity, precision, false negative rate and false discovery rate.

3.3 *Results*

Figure 4 contains an input image (*left*) and the number of nuclei discovered when the threshold parameter is varied from 130 to 200 (*center*). For the same input image as in Fig. 4 (*left*), Fig. 5 illustrates the output of the distance transform method for different threshold values. The plot in the right from Fig. 4 illustrates a comparison

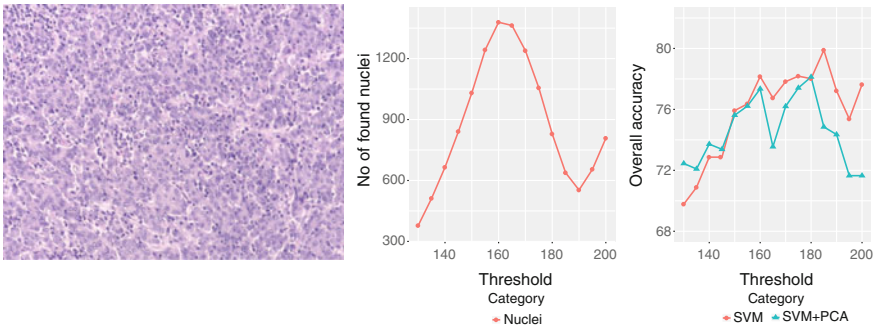

Fig. 4 The detected nuclei depend on the input threshold parameter before the distance transform method. The input (*left*) is a section of a histopathological image of G3 cancer, while the plot in the *middle* contains the number of nuclei there are detected when the input threshold parameter prior to the distance transform step is varied from 130 to 200. The plot on the *right* illustrates a comparison between the SVM applied to the complete data and the same classifier on the PCA-reduced data

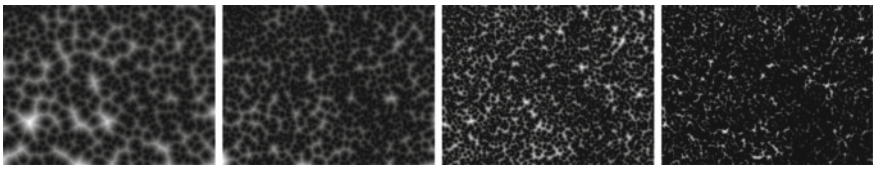

Fig. 5 The distance transform images when the threshold value is 130, 150, 170 and 190, respectively. The input image is the same from Fig. 4 (*left*)

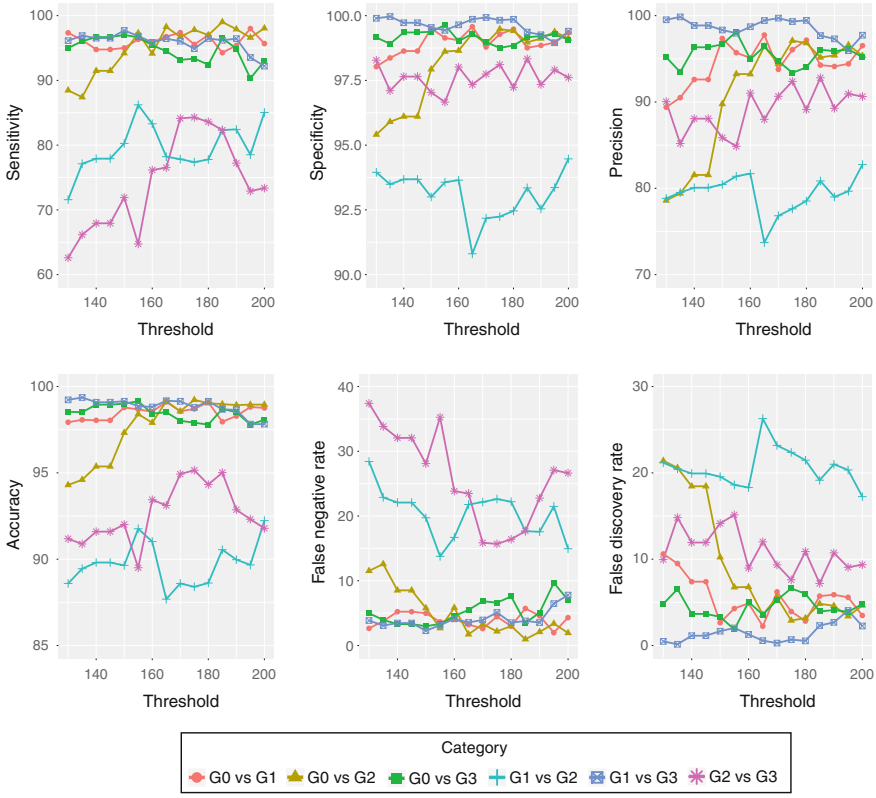

Fig. 6 Comparison between grades taken two by two: sensitivity, specificity, precision (*first line*), accuracy, false negative rate, false discovery rate (*second line*) after SVM classification

between the SVM applied on the entire numerical data, on the one hand, and on the principal components, on the other hand. As the results are generally superior in the case when the data is kept unaltered, the PCA is next left aside and Fig. 6 contains the detailed comparisons for each grade in turn in the direct SVM application only.

3.4 Observations and Discussion

Through its nature, the distance transform achieves a relatively good separation of nuclei that intersect or even marginally overlap. However, if two or more nuclei overlap at a great extent, they will be labeled as one object. As Fig. 4 (*right*) and Fig. 5 suggests, the higher the threshold value is (over 170), the less objects are detected. Although Fig. 5 indicates that more nuclei are detected as the threshold value increases, they overlap in the last image, conducting thus to a smaller number of found objects. Despite the general and natural opinion that the overlapping nuclei should be identified as separate entities in order to provide accurate data to the classifier, the best overall accuracy result (79.89 %) is obtained when the threshold is 185. In contrast to other studies, herein the focus does not lie into clinically assessing the automated identification capabilities, but instead a more objective measure is followed, that of choosing this proper threshold parameter for the detection of nuclei in correspondence to the prediction accuracy only.

The plots in Fig. 6 place the normal tissues (G0) very well as compared to the rest of the grades, especially when the threshold value is higher than 155. The distinction between G0 and the other grades represented a major concern for the current data set because the histopathological images for the normal tissue were all obtained by cutting pieces from larger images that belonged to G1 and mostly G2 and G3. This means that the same light conditions, the same amount of H&E and the exact same settings of the microscope were used when producing them. These results are very encouraging as regards the practical use of the proposed technique because, by removing even solely computer diagnosed G0 images, the pathologist can focus on the most difficult cases.

The weakest results are when comparing G1 versus G2 and G2 versus G3, as all the plots indicate. The specificity, precision and false discovery rate show that the hardest cases to be distinguished are between G1 and G2, as it actually occurs for the pathologists, as well [17].

A similar methodology where both nuclei and glands are taken into account for prostate cancer diagnosis, also with four grades to discriminate between, is presented in [17]. Although the measures used in the study are clearly described, the authors do not provide details about the manner of identifying the two types of components. The same SVM classifier is employed. The accuracies achieved between grades in [17] are between 76.9 and 92.8 %, while in the current study they are between 90.55 % (G1 vs. G2) and 98.98 % (G0 vs. G2) for threshold 185, in Fig. 6 (*first image, second line*). Naturally, the problem is not the same, hence the current image data set is made available [6] for future studies and direct comparison.

4 Conclusions and Future Work

A data set of 357 histopathological images is put forward, each having a resolution of 800×600 pixels, separated in 4, well balanced grades. Also, a diagnosis support methodology that uses morphological and topological features is proposed, which conducts to an overall accuracy of 79.89 %. The glands are discovered by applying a Gaussian filtering followed by a watershed algorithm, while the nuclei are found after a normalized box filter and distance transform are used.

The influence of a threshold parameter used for identifying the nuclei over the results indicates a surprising information: although in some pictures the discovered overlapping nuclei are merged in larger components, it is not only that the accuracy is not decreased, but some improvement is reached.

As the normal tissues are extracted from larger images that contained marginal separations between different grades of cancer and normal tissue, it is intended to add intensity-based features in a future work to see if this can boost the automatic diagnosis or, on the contrary, misleads the identification of the normal tissue from the other ones. Also, other topological information, like co-adjacency matrices could conduct to better results. Probably a gain in prediction could be obtained by fine tuning the parameter values of the methods involved in the identification of the main components.

Acknowledgments Present work was supported by the research grant no. 26/2014, code PN-II-PT-PCCA-2013-4-1153, entitled IMEDIATREAT—Intelligent Medical Information System for the Diagnosis and Monitoring of the Treatment of Patients with Colorectal Neoplasm—financed by the Romanian Ministry of National Education (MEN)—Research and the Executive Agency for Higher Education Research Development and Innovation Funding (UEFISCDI).

References

1. Mills, S.: Histology for Pathologists, 3rd edn. Lippincott Williams & Wilkins (2006)
2. Gorunescu, F., Belciug, S.: Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization. J. Biomed. Inf. **49**, 112–118 (2014)
3. Lee, H., Chen, Y.P.P.: Image based computer aided diagnosis system for cancer detection. Expert Syst. Appl. **42**(12), 5356–5365 (2015)
4. Gurcan, M., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009)
5. Demir, C.G., Kandemir, M., Tosun, A.B., Sokmensuer, C.: Automatic segmentation of colon glands using object-graphs. Med. Image. Anal. **14**(1), 1–12 (2010)
6. IMEDIATREAT Project. https://sites.google.com/site/imediatreat/
7. Sertel, O., Kong, J., Catalyurek, U.V., Lozanski, G., Saltz, J.H., Gurcan, M.N.: Histopathological image analysis using model-based intermediate representations and color texture: follicular Lymphoma grading. J. Sign. Process. Syst. **55**, 169–183 (2009)
8. He, L., Long, L.R., Antani, S., Thoma, G.R.: Histology image analysis for carcinoma detection and grading. Comput. Meth. Prog. Bio. **107**(3), 538–556 (2012)
9. Stoean, C., Stoean, R., Sandita, A., Mesina, C., Gruia, C.L., Ciobanu, D.: Evolutionary search for an accurate contour segmentation in histopathological images. In: ACM Genetic and Evolutionary Computation Conference, GECCO Companion, pp. 1491–1492. Spain, Madrid (2015)

10. Stoean, C., Stoean, R., Sandita, A., Mesina, C., Ciobanu, D., Gruia, C.L.: Investigation on parameter effect for semi-automatic contour detection in histopathological image processing. In: IEEE Post-Proceedings of the 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (in press). Timisoara, Romania (2015)
11. Beucher, S., Lantuej, C.: Use of watersheds in contour detection. In: International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation. Rennes, France (1979)
12. Kimmel, R., Kiryati, N., Bruckstein, A.M.: Sub-pixel distance maps and weighted distance transforms. J. Math. Imaging Vis. **6**, 223–233 (1996)
13. Irshad, H., Veillard, A., Roux, L., Racoceanu, D.: Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. IEEE Rev. Biomed. Eng. **7**, 97–114 (2014)
14. Kong, J., Cooper, L., Kurc, T., Brat, D., Saltz, J.: Towards building computerized image analysis framework for nucleus discrimination in microscopy images of diffuse glioma. In: IEEE 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6605–6608. Boston, MA, USA (2011)
15. Pang, Q., Yang, C., Fan, Y., Chen, Y.: Overlapped cell image segmentation based on distance transform. In: 6th World Congress on Intelligent Control and Automation, vol. 2, pp. 9858–9861. IEEE Press, Dalian (2006)
16. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: 10th European Conference on Machine Learning, pp. 137–142. Springer, London (1998)
17. Doyle, S., Hwang, M, Shah, K., Madabhushi, A., Feldman, M.D., Tomaszeweski, J.E.: Automated grading of prostate cancer using architectural and textural image features. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1284–1287. Washington, DC (2007)