

Numerical characterization of the images of prostate cancer for recognition of Gleason scale

Abstract. The paper presents the algorithm of numerical characterization of the histological structures existing in the images of the prostate cancer able to associate the image with the Gleason scale. These descriptors characterize the geometry of ducts by applying different measures, like the area defined in different way, perimeter, two types of diameters of the ducts (the short and long) as well as some relative coefficients relating different geometrical parameters to each other. We will analyze the numerical characterization of the discrimination abilities of different descriptors by applying Fisher measure and the application of the principal component analysis (PCA) to develop the most efficient feature set, able to associate the graphical image of the biopsy with a Gleason scale in an automatic way by applying Support Vector Machine as the classifier.

Streszczenie Praca przedstawia algorytm parametryzacji elementów struktur histologicznych w obrazach mikroskopowych prostaty która może znaleźć zastosowanie w automatycznym rozróżnianiu klas chorobowych raka prostaty opisanych skalą Gleasona. Obrazy struktur wydzielone z obrazu poddano procesowi parametryzacji, czyli opisowi cech przy użyciu deskryptorów numerycznych. Jakość poszczególnych deskryptorów zbadano przy zastosowaniu miary Fishera i transformacji PCA, których wyniki wskazują jednoznacznie na cechy mogące mieć znaczenie przy rozwiązywaniu zadania automatycznej klasyfikacji obrazów, przypisując ich wygląd do określonej skali Gleasona. Ostatni etap rozpoznania skali Gleasona jest realizowany przy użyciu klasyfikatora typu Support Vector Machine, który dokonuje przypisania danego obrazu do określonej skali Gleasona. (Algorytm parametryzacji elementów struktur histologicznych w obrazach mikroskopowych prostaty)

Keywords: mathematical morphology, Gleason grade, parameterization, image processing.

Słowa kluczowe: morfologia matematyczna, skala Gleasona, parametryzacja, przetwarzanie obrazów.

Introduction and medical basis

The paper will present the algorithm of numerical characterization of the histological structures existing in the images of the prostate cancer which provides compatibility with the Gleason scale [5]. The Gleason scale enables the likely aggressiveness of prostate cancer to be measured. Fig. 1 illustrates the graphical connection between the view of the histological structure and the Gleason scale [12]. The prostate cell is given the grade of 1 when it looks like the normal cell. The grade of 5 is given to the most abnormal and undifferentiated cancerous cells. Gleason grades from 2 to 4 are the intermediate grades [1-4].

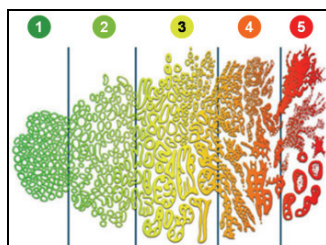


Fig. 1 The illustration of the Gleason scale on the basis of the histological structures of the ducts [12]

The basic task of the research is to create fully automatic classifier able to associate the prostate biopsy image with the Gleason grade of the prostate cancer. Such automatic system will do all steps automatically, starting from the image of the biopsy and finally pointing the level of the Gleason scale. In this way the developed system will be able to support the medical diagnosis of this type of cancer.

In our experiments we have analyzed the data corresponding to 32 images of biopsy corresponding to Gleason scales from 2 to 5. Among these cases there were nine images representing scale 2, seven of scale 3, eleven of scale 4 and five belonging to the scale 5.

Extraction of the duct structure

The first stage of the image processing is aimed on the extraction of the ducts existing in the image. We have solved this problem by applying mathematical morphology operations, such as erosion, dilation, opening, closing, hole

filling, etc. [6,7]. The detailed description of this procedure was presented in [13] and includes creation of binary image, recognition of the tissue contour, formation of ducts contour, filling these contours and finally separation of the individual ducts for further parameterization. Fig. 2 illustrates the most important steps of the ducts extraction.

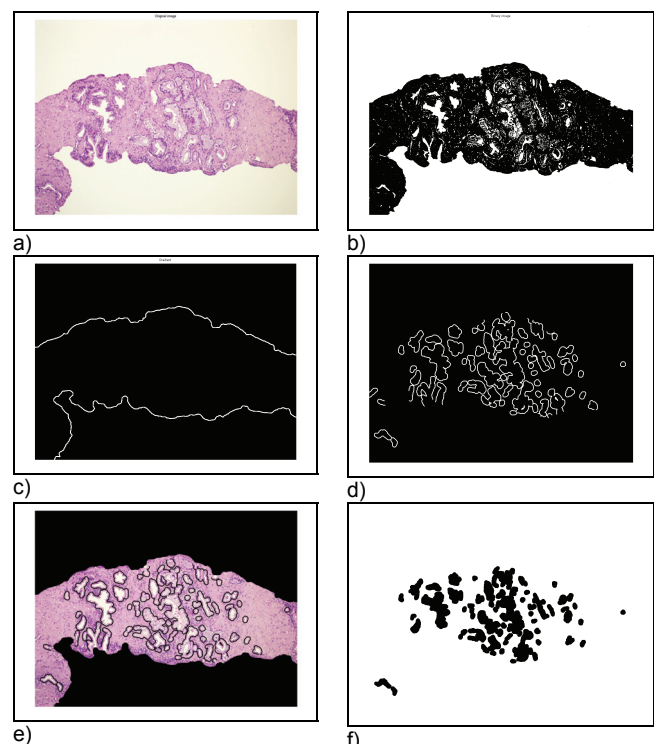


Fig. 2 The most important stages of the extraction algorithm: a) original image, b) binary image, c) tissue contour, d) ducts contours, e) visualization of the ducts on the tissue image, f) final result of segmentation

The ducts parameterization

The extracted ducts are subject to the numerical characterization by using geometrical descriptors. The values of these descriptors characterize the geometry of the

ducts by applying different measures, like the area (defined in different way), perimeter, two diameters of the ducts (the short and long) as well as some relative coefficients relating different parameters to each other. In this way the whole set of geometrical parameters describing the ducts existing in the image of the biopsy are created. In particular the following descriptors have been directly determined: the number of ducts – N , area of the ducts – A , convex area – A_c , sum of the areas of ducts – A_s , area of the tissue – A_{tk} , long diameter – D_l , short diameter – D_s , distance between the neighboring ducts – D_n , perimeter – P , convex perimeter – P_c . Besides these parameters we have defined some relative measures. They include [13]

- elliptic factor

$$(1) \quad F_e = \frac{D_s}{D_l}$$

- circularity factor

$$(2) \quad F_c = \frac{4\pi A}{P^2}$$

- factor of convex circularity

$$(3) \quad F_{cc} = \frac{4\pi A_c}{P_c^2}$$

- homogeneity factor

$$(4) \quad F_h = \frac{P^2}{A}$$

- convex homogeneity factor

$$(5) \quad F_{ch} = \frac{P_c^2}{A_c}$$

- corrugation factor

$$(6) \quad F_{cr} = \frac{P_c}{P}$$

- factor of ruggedness

$$(7) \quad F_r = \frac{A}{A_c}$$

These parameters describing the geometry of the ducts extracted from the image are potential diagnostic features that can be used by the neural classifier to recognize different patterns of the histological structures. For the purpose of further processing we have applied the statistical normalization of the data according to the relation

$$(8) \quad f_i = \frac{f_i - \text{mean}(f)}{\text{std}(f)}$$

in which f_i is the i th value of feature f , while mean and standard deviation (std) are calculated for all values of f available in the set of data.

The feature selection

In the next step we have to characterize all extracted ducts by the mentioned geometrical parameters and then use the chosen measure of quality to select the most discriminative features. In our study we have used the well known Fisher quality measure $S_{AB}(f)$ of the feature f

$$(9) \quad S_{AB}(f) = \frac{|c_A(f) - c_B(f)|}{\sigma_A(f) + \sigma_B(f)}$$

In this definition c_A , c_B and σ_A , σ_B are the mean values and standard deviations of the feature f in the class A and B , respectively. The higher is the value of $S_{AB}(f)$ the better is the ability of the feature f to recognize between the classes A and B .

Table 1 presents the values of this Fisher measure [12] for each of the candidate feature at recognition of images representing four investigated classes (nine representatives

of scale 2, seven of scale 3, eleven of scale 4 and five of scale 5). As we can see the ability of each feature to recognize two classes is changing for different combination of classes. We have to select the feature which is the best for all classes. Therefore in the last column we have added the summed discriminative power for recognition of all pairs of classes. They have been arranged according to this sum in an increasing order. To the best features belong: N , F_e , F_{ch} and A_c . On the other side the least discriminative are F_{cer} and F_r . It is difficult to decide in advance where is the boundary of the best features. It is even more challenging since at cooperative performance of all features their discriminative ability may change.

Table 1 Fisher measure of the geometrical parameters characterizing prostate cancer of four Gleason scales (four classes)

Descriptor	S23	S24	S25	S34	S35	S45	SUM
F_{ce}	0,08	0,09	0,10	0,00	0,02	0,02	0,30
F_r	0,00	0,13	0,15	0,11	0,13	0,00	0,52
F_{cc}	0,54	0,04	0,03	0,51	0,49	0,00	1,61
d_l	0,32	0,13	0,33	0,36	0,51	0,17	1,81
P	0,37	0,47	0,59	0,14	0,32	0,18	2,07
A_{tk}	0,24	0,41	0,23	0,65	0,43	0,12	2,08
d_n	0,11	0,46	0,59	0,34	0,48	0,17	2,14
A_s	0,27	0,48	0,33	0,55	0,33	0,22	2,18
d_s	0,80	0,52	0,63	0,15	0,07	0,07	2,24
F_c	0,91	0,06	0,00	0,64	0,59	0,04	2,24
F_h	0,75	0,45	0,64	0,23	0,09	0,15	2,30
P_c	0,53	0,54	0,75	0,12	0,37	0,21	2,52
A	0,16	0,44	0,68	0,55	0,78	0,22	2,83
A_c	0,16	0,62	0,84	0,72	0,92	0,23	3,48
F_{ch}	1,29	0,49	0,81	0,65	0,37	0,27	3,88
F_e	0,64	0,82	1,31	0,24	0,61	0,29	3,91
N	0,07	0,62	1,27	0,44	0,94	0,60	3,94

To solve the problem of selection we have applied principal component analysis (PCA) of the data. PCA is a method of the linear transformation of the original data set represented by vector samples \mathbf{x} into a new set of vector samples \mathbf{y} with the reduced dimensions, where $\mathbf{y}=\mathbf{W}\mathbf{x}$ with \mathbf{W} the PCA transformation matrix [8]. Presenting the results of PCA on the plane allows the user to assess the recognition capability of the presented feature set in a visual way. Fig. 3 presents the distribution of the eigenvalues of the autocovariance matrix \mathbf{R}_{xx} corresponding to all 17 features. As it is seen this is the uniform exponential distribution and once again it is very hard to establish the border between the most and least important features for automatic recognition of the Gleason scale.

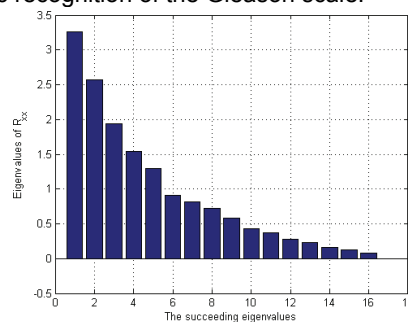


Fig. 3 The distribution of the succeeding eigenvalues of the autocovariance matrix \mathbf{R}_{xx}

To find the best set of features we have applied the graphical presentation of the multidimensional data mapped on two most important principal components PCA_1 and PCA_2 at application of different number of the features. The criterion of optimality is the best separation of data belonging to different classes (Gleason scales). Fig. 4 illustrates the distribution of data belonging to four classes at application of all 17 original features. The sign of \times represents Gleason scale 2, the star – scale 3, the sign of triangle down – scale 4 and pentagram – scale 5.

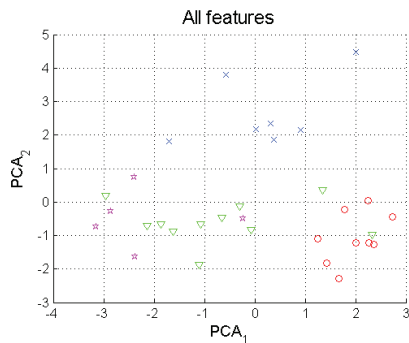


Fig. 4 The dataset projection of the data represented by full set of features on the 2 principal components

To find the best set of features we have mapped the original data on the PCA two-dimensional system at assumption of different numbers of features, and assessing the data distribution in a visual way. We got the best discrimination of classes at representation of data by 11 best features. Fig. 5 presents this distribution of data at application of only 11 best features in the vector x . As it is seen the representation of data by 11 features seems to be much better in comparison to the case of all features (smaller number of interlacing samples of different classes of data).

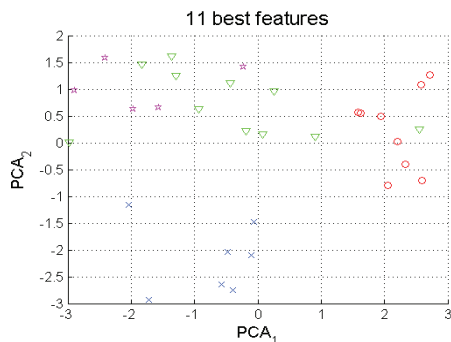


Fig. 5 The dataset projection of the data represented by 11 best features on the 2 principal components

The dataset presented in both figures contains images coming from 32 patients of different stages of development of the prostate cancer (Gleason scale from 2 to 5).

The SVM classification of data

The final stage of the procedure is to classify the data into four predefined classes. Since the number of data is scarce, the best results of recognition are achieved at application of Support Vector Machine as the classifier, known as the excellent tool of good generalization ability for classification problems, even at small representation of data [14]. To deal with many classes we have applied one-against-one strategy [8].

The Gaussian kernel SVM classifiers have been found the best in our application. The optimal values of regularization parameter C and hyperparameter σ of the Gaussian function were determined for each pair of classes independently, after additional series of learning experiments through the use of the validation test sets. Many different values of C and σ combined together in the learning process have been used in the learning process and their optimal values are those for which the classification error on the validation data set was the smallest one ($C=1000$, $\sigma=0.05$). The SVM networks were trained using modified Platt algorithm [8,10]. To get the most objective results at small number of data we have to apply the leave-one-out method. 31 data points were used to train the SVM classifier and the last one used to test it.

This process was repeated 32 times and the final score is the mean of all trials. At application of all 17 features we have got 5 wrong recognized samples (the relative error equal 15,6%). Reducing the number of features to the best 11 parameters the number of misclassified samples was reduced to 1 (the relative error equal 3%). This confirms the significance of the selection stage of the features.

Conclusions

The paper has presented the computerized system for the automatic analysis of the biopsy image of the prostate cancer. The recognition of the Gleason scale is done by extracting the histological duct structure, parameterization of the extracted ducts and using the selected number of these parameters as the diagnostic features in the final recognition step performed by applying SVM. The aim of the SVM classifier, doing this step is to associate the duct structure of the image with the Gleason scale. The important advantage of the solution is the automation of this very difficult work, not possible to be done manually by the human expert. The developed system was applied to the analysis of 32 medical images corresponding to different stages of the prostate cancer (Gleason scale) and the results are very encouraging.

REFERENCES

- [1] Cieśliński P., Dadej R., Kwias Z., Rak stercza. *Współczesna Onkologia*, 2002, vol. 6, 2, 108-116
- [2] Duncan W., Recent results in cancer research #78: Prostate cancer. Heidelberg: *Springer Verlag*, 1981
- [3] Grayhack J., Keeler T., Kozłowski J., Carcinoma of the prostate. *Cancer*, 2006, vol. 60, pp.589 - 601
- [4] Peeling W. B., Algaba F. Update on urology prostate cancer. *EJSO*, 1996, vol. 22, no1, pp. 102-106
- [5] Gleason D.F., Mellinger G.T., Prediction of prognosis for prostatic adenocarcinoma by histological grading and clinical staging. *Journal of Urology*, 1974, vol. 111, pp. 58-64
- [6] Soile P., Morphological image analysis, principles and applications. Berlin: *Springer Verlag*, 2003.
- [7] Tadeusiewicz R., Korohoda P., Komputerowa analiza i przetwarzanie obrazów. Kraków: *Wyd. Post. Telekom.*, 1997.
- [8] Osowski S., Sieci neuronowe do przetwarzania informacji. Warszawa: *Oficyna Wydawnicza PW*, 2006
- [9] Kruk M., Automatyczny system rozpoznawania komórek na podstawie obrazu mikroskopowego wybranej tkanki ludzkiej dla potrzeb diagnostyki medycznej, *Ph.D. PW*, 2008
- [10] Matlab user manual, MathWorks, 2007
- [11] Cytowski J., Gielecki J., Gola A., Cyfrowe przetwarzanie obrazów medycznych., Warszawa: *Exit*, 2008
- [12] <http://talkaboutprostatecancer.files.wordpress.com>
- [13] Kruk M., Osowski S., Koktyś R., Segmentacja i parametryzacja struktur histologicznych w obrazach mikroskopowych prostaty dla oceny skali Gleasona. *Przegląd Elektrotechniczny*, 2010, vol. 86, No 5, pp. 5-9
- [14] Schölkopf B., Smola A., Learning with kernels, Cambridge, *MIT Press, MA*. 2002

Authors: dr inż. Michał Kruk, Warsaw University of Life Sciences, Faculty of Applied Informatics and Mathematics Email: krukm@iem.pw.edu.pl;
 prof. dr hab. inż. Stanisław Osowski, Warsaw University of Technology, Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Military University of Technology, Institute of Electronic Systems, Email: sto@iem.pw.edu.pl.
 med. dr Robert Koktyś, Department of Pathology, Military Institute of Medicine, rkoktyś@poczta.onet.pl.