The final publication is available at IEEE Xplore® via https://doi.org/10.1109/CPEE.2017.8093057. Cite as:
G. Wieczorek, B. Świderski, L. J. Chmielewski, M. Kruk, and A. Orłowski. Choice of distance function in the segmentation of regions of interest in microscopic images of breast tissues. In Proc. 18th Int. Conf. Computational Problems of Electrical Engineering CPEE 2017, pages 14, Kutná Hora, Czech Republic, 11-13 Sep 2017. IEEE, 2017. doi:10.1109/CPEE.2017.8093057.

Choice of Distance Function in the Segmentation of Regions of Interest in Microscopic Images of Breast Tissues

Grzegorz Wieczorek, Bartosz Świderski, Leszek J Chmielewski, Michał Kruk and Arkadiusz Orłowski
Faculty of Applied Informatics and Mathematics – WZIM
Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-775 Warsaw, Poland
Email: {grzegorz_wieczorek,michal_kruk}@sggw.pl
Internet: www.wzim.sggw.pl

***Abstract*—Classification of milk duct carcinoma in the scans of diagnostic specimens is an important medical problem. Before the classification is performed, the regions of milk ducts which will be the regions of interest (ROI) should be detected. One of the approaches to such detection is to segment the image into ROIs and the remaining regions. The segmentation by clusterization with the classical K-means method was proposed in the literature. A pixel together with its square neighborhood was considered as the object. Sorted image intensities in the neighborhood with extreme values omitted were used as features, with the Euclidean distance between the objects. In this paper we investigate new distance functions: cosine distance, city block and correlation distance, in the same setting. The cosine function was found to be the best, giving smaller average error, as well as smaller scatter measure, with respect to the Euclidean function. The mean errors for the cosine, Euclidean, city block and correlation functions were 17%, 25%, 39% and 89%, respectively.**

***Index Terms*—ROI detection, clusterization, milk ducts, distance function, cosine, city block, correlation, K-means**

I. Medical background

At present, the ductal carcinoma in situ (DCIS) constitutes over 20% of disclosed neoplastic changes of the mammary gland. The only symptom of the disease is usually an incorrect result of the mammographic examination, and in $1/3$ of detected cases an invasive carcinoma can develop. The ductal carcinoma sometimes manifests itself with a palpable tumor or a pathological discharge from the nipple in the course of the Paget disease of the breast [1]. Over 90% of cases remains asymptomatic at the time of disclosure. Usually, a biopsy is prescribed due to an incorrect result of a radiologic examination [2]. The final diagnosis is made on the basis of the assessment of tissue sections sampled during the biopsy of the breast. In the past it was necessary to extract the whole suspected fragment of the gland, but the contemporary methods made it possible to collect a fragment with the biopsy needle.

The following phases of advancement of the ductal carcinoma can be distinguished (cf. [3] → Symptoms & Diagnosis → Types of Breast Cancer → DCIS → Diagnosis of DCIS):

- **Ductal hyperplasia** The overgrowth of cells: the number of cells is larger than in a normal mammary duct.
- **Atypical ductal hyperplasia** The number of cells is too large (hyperplasia) and the appearance of some of them is atypical.
- **Ductal carcinoma in situ (DCIS)** There is hyperplasia and the cells have the features of cancer, but they do not pass the borders of the duct (Figs. 1a and 1b).
- **DCIS-MI (DCIS with microinvasion)** A limited number of cancer cells slightly infiltrate the duct wall.
- **Invasive ductal carcinoma** The cancer cells infiltrate the tissues beyond the mammary duct. In this phase the carcinoma is not a DCIS, but it is denoted as invasive ductal carcinoma (IDC), or more precisely, invasive ductal carcinoma of no special type (NST). This is the most common type of breast cancer (Figs. 1c and 1d show two types of this carcinoma).

The detailed diagnostic criteria are available in specialist literature and in the recommendations of the World Health Organization [6].

II. Objective

The analysis of the types of carcinoma should be performed only in the regions of the images of the tissue sections which contain the cells suspected of being cancerous. Such analysis can be time-consuming and it would be impractical to perform

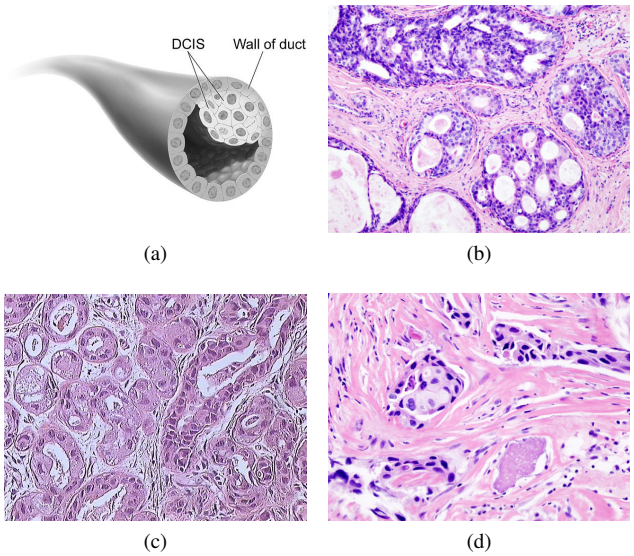

Fig. 1. Phases of advancement of the ductal carcinoma. (a) Schematic image of DCIS. By Don Bliss – National Cancer Institute. Image released by the National Cancer Institute, an agency part of the National Institutes of Health, with the ID 4353. Public Domain, from [4]. (b) Histopathologic image of DCIS, hematoxylin and eosin stain. Public Domain, from [4]. (c) Invasive ductal carcinoma, hematoxylin and eosin stain. Public Domain, from [5]. (d) Invasive ductal carcinoma, a scirrhous type of growth, hematoxylin and eosin stain. Public Domain, from [5].

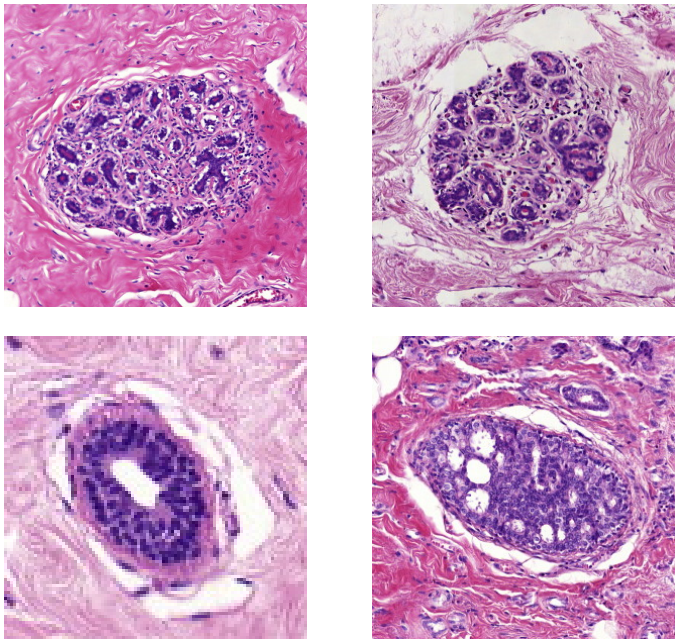

Fig. 2. Examples of original images used in experiments. Sizes of windows shown are different.

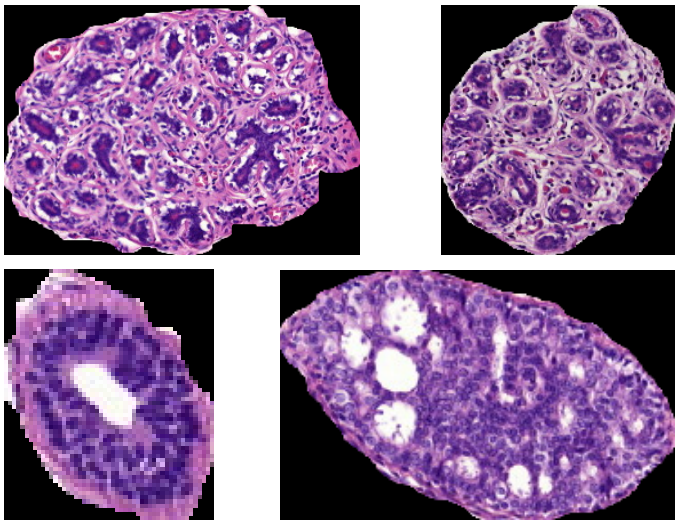

Fig. 3. Examples of ground truth regions prepared by the expert for images of Fig. 2.

it in the whole images. The cells of the milk ducts in the images normally analyzed by human observers differ to some extent from other cells, which is the results of staining. Only the regions which contain such cells should be analyzed further and these are the regions of interest (ROI) which are sought in this study. As the ground truth GT, the regions of interest marked by the expert will be used. The objective is to find the ROIs which fit the ground truth in the best way. Ideally, the ROIs found should be identical with the GT, but inevitably some false positive and false negative errors can arise.

In our previous paper [7] we have proposed and tested the method for finding the ROIs with the use of regional features with some robustness properties and with K-means clustering. The classical Euclidean distance function was used. In this paper we shall study three other distance functions: cosine, city block and correlation distance in order to check whether the error of ROI detection can be reduced.

III. Materials

The images for analysis were provided by the Military Institute of Medicine, Division of Pathomorphology (Wojskowy Instytut Medyczny, Zakład Patomorfologii) in Warsaw. They were collected by scanning the specimens from biopsy with a high-resolution scanner 3D Histech. We had 250 samples from 25 patients. In Fig. 2 selected images from the set used in the experiments are shown.

The ground truth images were produced by manually marking the regions of interest (see Fig. 3 for the results for objects from Fig. 2). It should be noted that the ground truth has been prepared by the expert in the form of a selection of the milk duct region in the image. In this study the accuracy attained by the human expert was not tested. This could be done by comparing the selections made by a number of experts and by the same experts at different times. This will be the subject of another study. At present, the available data will be used.

IV. The proposed method

Let us briefly present the method already described in [7]. We shall use the same example as in that paper but the description will be shortened. The parameters found in [7] as giving the best results will be assumed throughout.

A. Segmentation by clustering

Optimally, the region of interest (ROI) should contain all and only the pixels belonging to the cells of the milk ducts together with the neoplastic cells, including the cells possibly infiltrating the surrounding tissues. It is admissible that the interior of the milk ducts is also included in these regions.

The pixels of the image were clustered into sets, or clusters, with the assumption that only one of these clusters will be the the region of interest sought. The clustering method was the classical K-means method [8]. It is not vital that there are only two clusters – the ROI and the remaining region, so the number of clusters is not limited to two and various numbers were be tested. As a result, $K = 4$ was chosen, as giving the smallest segmentation error. It should be stressed that there is no classification process, hence, no teaching in the algorithm. On the contrary, the image is segmented. The only adaptation process is the choice of the parameters of the algorithm.

A method of selecting the cluster to be considered as the ROI is in preparation and will be published separately. At present the proper cluster is selected manually. It should be stressed, however, that this was a single cluster, not a sum or any other set-theoretic function of two or more clusters.

B. Features

Basically, the object considered in the clustering process is a pixel. However, to make it possible to extract more features for such an object it is a common practice to consider a pixel together with its neighborhood. A square neighborhood to be used here will be characterized by its half-width, here referred to as the *order* n; hence, the neighborhood of order n has size $(2n+1)\times(2n+1)$. Each pixel has three color components. The RGB color space, directly as received from the scanner, will be used here, although other spaces could appear to function better in this setting.

The data from the neighborhood are used in the following way. The intensities in each color channel are sorted non-decreasingly, and the values for the colors are concatenated, so the feature vector $\mathbf{x}$ of length $3(N+1)$ is formed: $\mathbf{x} = [R_0, R_1, \ldots, R_N, G_0, G_1, \ldots, G_N, B_0, B_1, \ldots, B_N]$, where $N = (2n+1)(2n+1) - 1$. Hence, $\mathbf{x}$ has $3(2n+1)(2n+1)$ coordinates.

Sorting of intensities improves the insensitivity of the algorithm to outliers in a simple way: P extreme values of pixel intensities are excluded from consideration. Then, $\mathbf{x}$ has $3[(2n+1)(2n+1) - 2P]$ coordinates. This is a direct and simple application of order statistics [9], where the smallest and largest statistics are postponed as susceptible to extremity.

Let us consider an example fragment of an image shown in Fig. 4. The pixel marked with grey background, having coordinates row $=$ 3, column $=$ 4, is represented in the neighborhood of order $n = 2$, marked with bold font in Fig. 4,

TABLE I
PARAMETERS OF THE ERROR MEASURES OBTAINED FOR THE FOUR DISTANCE FUNCTIONS.

	cosine	K**-means**	**city block**	**correlation**
mean	16,88	25,37	38,94	88,57
standard deviation	13,29	26,42	59,90	117,67
min	6,11	6,14	9,24	7,46
max	60,05	124,39	259,71	371,34
median	10,73	15,98	16,41	20,66

with the following vector:

$$\mathbf{x}_{3,4} = [\cancel{30},\cancel{30},53,60,65,67,81,89,108,117,149,155,157,162,$$
$$164,165,169,169,173,181,197,213,215,\cancel{222},\cancel{240},$$
$$\cancel{27},\cancel{28},35,38,49,55,89,93,100,103,114,124,133,133,$$
$$134,135,150,160,178,189,195,220,238,\cancel{249},\cancel{254},$$
$$\cancel{42},\cancel{50},55,69,83,84,96,100,101,101,108,112,113,174,$$
$$180,188,196,196,202,206,210,213,220,\cancel{227},\cancel{243}] .$$

In this formula, $P = 2$ extreme values for each color have been crossed out to mark that they are not be considered.

For pixels close to the border for which a part of neighborhood is missing, and in the case of sorting the intensities, the missing part of the neighborhood is complemented by doubling the necessary number of borderline verses and columns. According to [7] the best result was obtained for the order $n = 10$ and with $P = 2$ extreme values neglected.

C. Measure of error

The ROI images obtained as the result of selecting the proper cluster were compared to the ground truth ROI images by taking the modulus of their pixel-wise difference. Both these images are binary, so the difference image is also binary, with pixels outside (inside) both ROIs equal to zero (black) and pixels in which the ROIs differ equal to one (white). The example difference images are shown in Fig. 5. As the measure of segmentation error the number of white pixels in the difference image divided by the number of white pixels in the respective ground truth image is used.

V. RESULTS

In the previous paper [7] the Euclidean distance between the vectors $\mathbf{x}$ was used. Here we shall apply three more distance functions: cosine, city block and correlation, to check whether the way the distance is measured has an influence on the results of ROI detection. The parameters of the results obtained for the four distance functions are shown in Fig. 6, and the detailed values are given in Table I.

The errors received can be ordered according to the ascending mean error in the following way: cosine distance, Euclidean, city block, correlation distance. In this respect, the cosine distance function appeared better than the previously used Euclidean distance, with mean error 17% versus 25%. It makes the choice easier that all the other parameters, namely, minimum, maximum, standard deviation and median of the

48	108	53	117	215	222
73	164	181	169	213	67
23	165	60	197	65	81
147	173	30	89	157	30
174	162	155	169	149	240
139	241	115	106	138	165

122	133	103	238	134	189
163	254	114	249	135	133
139	55	93	49	220	89
165	27	195	35	124	38
139	28	160	178	100	150
184	16	197	24	171	67

11	188	84	227	96	112
193	101	108	100	55	213
62	174	69	196	202	196
113	180	50	101	243	42
176	113	210	206	83	220
91	5	110	193	171	253

Fig. 4. Example of a neighborhood of order $n = 2$ of the pixel in row $= 3$, column $= 4$, marked with grey background, in three color channels. The pixels belonging to the neighborhood written in bold font.

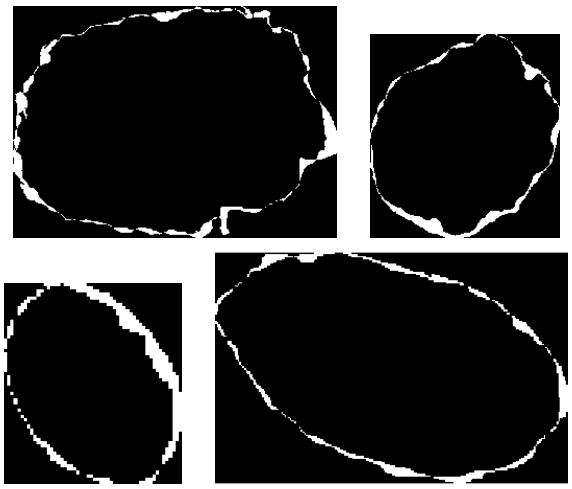

Fig. 5. Examples of error images of the ROIs from Fig. 3 found with the method.

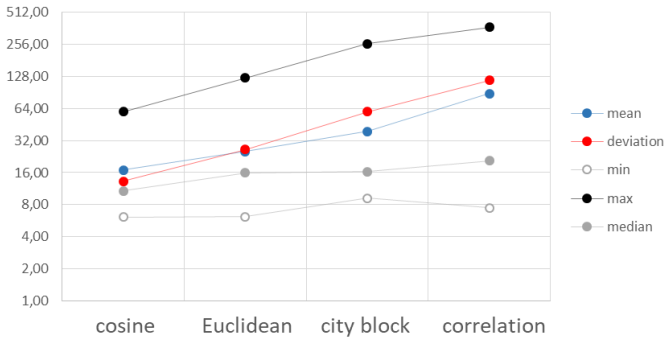

Fig. 6. Parameters of the error measure obtained for the four distance functions. Lines between points have no significance; they only indicate the relation between the points.

error nearly precisely follow the same tendency, with the values for the cosine distance being consistently the smallest.

VI. Conclusion

The scans of the images of specimens of the ductal carcinoma were analyzed and the regions of interest for future classification of potentially cancerous tissues was performed. This was done by segmentation in the way of clusterization with the K-means method. Pixels were clustered, with image intensities in the neighborhood with extreme values omitted used as features. In place of the Euclidean distance used in the previous paper, the cosine distance, city block and correlation distance were used. The mean errors of segmentation with these functions were 17%, 39% and 89%, respectively, while for the Euclidean distance used previously it was 25%. The cosine function was found the best also if the scatter measures were considered. This indicates that the choice of the distance measure has a significant influence on the segmentation error and can be used as a parameter in optimizing the result.

References

[1] C.-Y. Chen, L.-M. Sun, and B. O. Anderson, "Paget disease of the breast: Changing patterns of incidence, clinical presentation, and treatment in the U.S." *Cancer*, vol. 107, no. 7, pp. 1448–1458, 2006, doi:10.1002/cncr.22137.

[2] D. D. Dershaw, A. Abramson, and D. W. Kinne, "Ductal carcinoma in situ: mammographic findings and clinical implications," *Radiology*, vol. 170, no. 2, pp. 411–415, 1989, doi:10.1148/radiology.170.2.2536185.

[3] The Breastcancer.org Team, "Breastcancer.org," 2017, [Online; accessed 28 Jan 2017]. [Online]. Available: http://www.breastcancer.org

[4] Wikipedia, "Ductal carcinoma in situ — Wikipedia, The Free Encyclopedia," 2017, https://en.wikipedia.org/w/index.php?title=Ductal_carcinoma_in_situ&oldid=761575543 [Online; accessed 28 Jan 2017]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Ductal_carcinoma_in_situ&oldid=761575543

[5] ——, "Invasive carcinoma of no special type — Wikipedia, The Free Encyclopedia," 2016, https://en.wikipedia.org/w/index.php?title=Invasive_carcinoma_of_no_special_type&oldid=723403683 [Online; accessed 28 Jan 2017]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Invasive_carcinoma_of_no_special_type&oldid=723403683

[6] H.-P. Sinn and H. Kreipe, "A brief overview of the WHO classification of breast tumors, 4th Edition, focusing on issues and updates from the 3rd Edition," *Breast Care*, vol. 8, no. 2, pp. 149–154, 2013, doi:10.1159/000350774.

[7] B. Świderski, M. Kruk, S. Osowski, G. Wieczorek, J. Kurek, L. J. Chmielewski, and A. Orłowski, "Milk duct segmentation in microscopic HE images of breast cancer tissues," in *Proc. 21st Int. Conf. on Circuits, Systems, Communications and Computers CSCC 2017*, ser. MATEC Web of Conferences, N. Mastorakis, V. Mladenov, and A. Bulucea, Eds., vol. 125. Agia Pelagia Beach, Crete Island, Greece: EDP Sciences, 14-17 Jul 2017, p. 04013, doi:10.1051/matecconf/201712504013.

[8] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010, doi:10.1016/j.patrec.2009.09.011.

[9] M. Ahsanullah, V. B. Nevzorov, and M. Shakil, *An Introduction to Order Statistics*, ser. Atlantis Studies in Probability and Statistics. Amsterdam-Paris-Beijing: Springer, in cooperation with Atlantis Press, 2013, vol. 3, doi:10.2991/978-94-91216-83-1.