Milk duct segmentation in microscopic HE images of breast cancer tissues

Bartosz Świderski[1], *Michał* Kruk[1], *Stanisław* Osowski[2,3], *Grzegorz* Wieczorek[1], *Jarosław* Kurek[1], *Leszek J.* Chmielewski[1], and *Arkadiusz* Orłowski[1]

[1] Faculty of Applied Informatics & Mathematics, Warsaw University of Life Sciences, Warsaw, Poland
[2] Faculty of Electrical Engineering, Warsaw University of Technology, Warsaw, Poland
[3] Faculty of Electronics, Military University of Technology, Warsaw, Poland

Abstract. The aim of the paper is to recognize and extract the milk duct in haematoxylin and eosin (HE) stained breast cancer tissues. The paper presents the modified K-means approach to segmentation of the milk duct in HE stained images. Instead of using single pixels we propose to consider the defined region of pixels in the process. Thanks to such modification more accurate extraction of the milk ducts has been achieved. To compare the results in a numerical way the GT images prepared by the medical expert have been subtracted from the corresponding images created by the segmentation methods. The numerical experiments performed for many preparations have confirmed the superiority of such approach. The proposed method has allowed reducing significantly the error of duct segmentation in comparison to the classical K-means approaches. The results show, that our method is superior to the standard K-means and to the K-means preceded by averaging or Gaussian filtration at different size of filtration mask.

1. Introduction – medical background

Ductal carcinoma in situ (DCiS) belongs to the most frequently appearing type of non-invasive breast cancer [1,2]. It is "non-invasive" because is limited to the milk duct and is regarded as not life-threatening. However, DCiS increases the risk of developing an invasive breast cancer later on (around 1/3 od DCiS are ended in an invasive cancer). More than 90% of cases are without any visible symptoms in normal life.

The first signs of the illness can be recognized in the mammogram. After noting them the biopsy is usually made and the pathologist analyzes the piece of breast tissue and reports back on the type and grade of the DCIS, describing how abnormal the cells look when compared with the normal breast cells.

The image shown in Fig.1 [4] presents possible types of findings, starting from the normal cells and ending on the invasive ductal cancer. They include normal cells, ductal hyperplasia (too many cells present), atypical ductal hyperplasia (too many cells starting to take on an abnormal appearance), ductal carcinoma in situ (too many cells but still confined to the inside of the duct), DCiS with micro-invasion (few of the cancer cells breaking through the wall of the duct) and finally invasive ductal cancer (many cancer cells broken beyond the breast duct transferring DCiS into an invasive ductal carcinoma,).

Automatic recognition of the milk ducts is an important task in computerized approach to the problem. The aim is to separate individual milk ducts existing in the preparation. On the basis of many extracted ducts the medical expert can diagnose the particular case. Our task in this presentation is to develop computerized system which is able to extract the individual ducts from the whole image with the highest precision comparable to the precision of human expert.

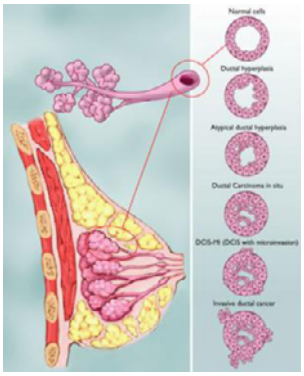


Fig. 1 Different phases of development of the DCiS. The images on the right represent the views of milk ducts in different stages of cancer advancement [4].

2. Problem statement

The aim of the paper is to recognize and extract the milk duct in haematoxylin and eosin (HE) stained breast cancer tissues. They form the region of interest in our analysis. An example of the analyzed image is presented in Fig. 2. It depicts the cross area of many milk ducts of different sizes and shapes. The arrows point to two chosen ducts. The group of ducts are located in the background of the tissue.

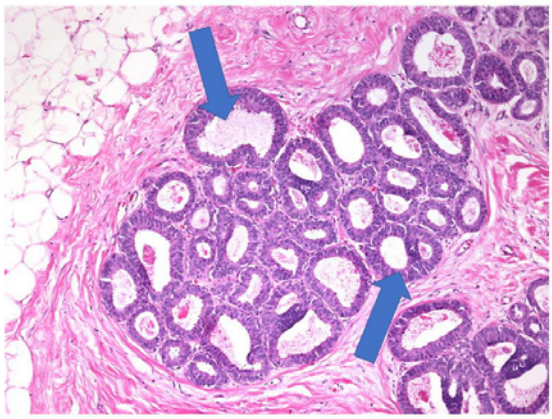


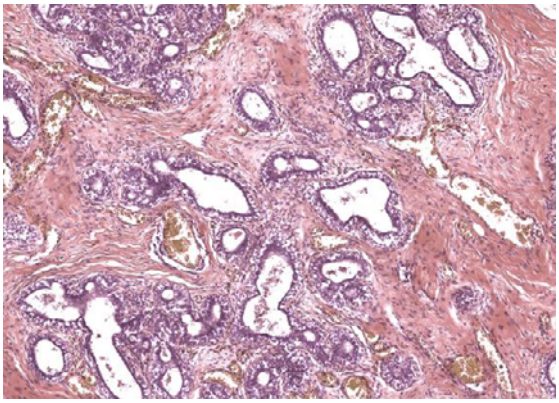
Fig. 2 An example of the image subject to the analysis in the paper. Many milk ducts of different size and shape, which should be extracted, are present in the image.

The most important factor in medical decision taking is precise determination of the shape, size, degree of filling the inside, width of the walls and channel patency. On the basis of values of these parameters the decision of the advancement of the breast cancer is taken. Therefore, the task of extraction of all milk ducts should be done as precise as possible.

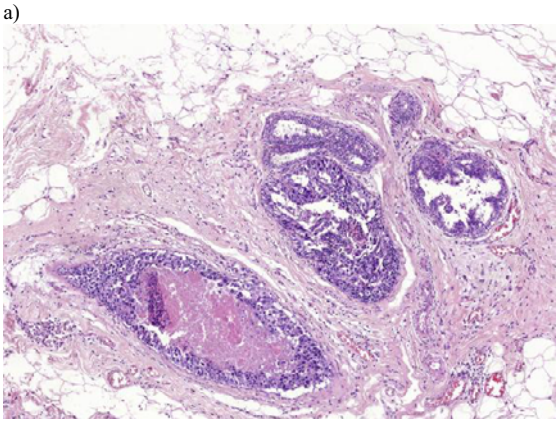
The popular method of doing it is the K-means algorithm [3,5,6]. However, its direct application to this problem is not efficient, since it results in a lot of artifacts in the background. The remedy to this problem is application of the morphological operations, which allow removing the residual noise in the resulting image. This process has some negative aspects - changing the shape and size of the milk duct. Our work presents some modification of K-means which eliminates the need for additional image processing. Thanks to this we are able to achieve better accuracy in the milk duct reconstruction. This will be confirmed by comparing our results with the ground truth (GT) pattern pointed manually by the medical expert for the same images.

3. Materials

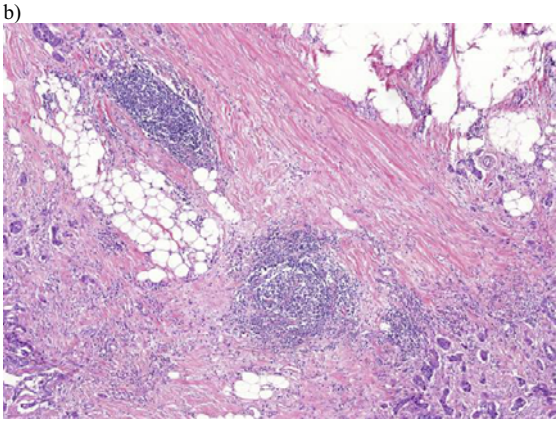
The HE images of the breast cancer tissues, which are subject to analysis, have been prepared in the Military Medical Institute in Warsaw. The biopsy material has been acquired from the patients suffering from breast cancer of different grade and then prepared using HE technology. The preparations have been scanned using high definition 3D Histech scanner and obtained images stored in the data base. The size of each image was equal 55808×82688. Fig. 3 presents some examples of images representing three grades of cancer advancement.



a)



b)



c)

Fig. 3 The examples of HE images of breast cancer tissues representing three different grades: a) ductual hyperplasia, b) ductual carcinoma in situ (DCIS), c) invasive ductual cancer.

4. Methods

The aim of clusterization is to separate the analyzed image into clusters representing the milk ducts existing in the image. Classical K-means operates on individual pixels in RGB channels, searching for the nearby pixels of similar intensity [5,6]. In our proposition we extend the representation of individual pixel to the region around it, including N neighboring pixels. In the case of RGB image each pixel will be represented by the

vector of the length 3N, **x**=[R_0, R_1, …, R_N, G_0, G_1, …, G_N, B_0, B_1, …, B_N], where the elements within R, G an B family are arranged in a non-decreasing order of pixel intensity values. An example of vector creation is shown below. Let us assume the RGB representation of image in the following matrix form.

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

We will create vector representation for the marked RGB pixels of the intensity levels equal 197, 49 and 196, respectively. Assume N=25, which means that each pixel of the image has 2 neighbors from left and right as well as from the top and bottom. The vector representation of the marked pixels neighborhood region in RGB channels arranged in a non-decreasing order will look as following.

$\mathbf{X}_1$=
[*30*,30,53,60,65,67,81,89,108,117,149,155,157,162,164,165,169,169,
173,181,197,213,215,222,240,
27,28,35,38,49,55,89,93,100,103,114,124,133,133,134,135,150,160,
178,189,195,220,238,249,254,
42,50,55,69,83,84,96,100,101,101,108,112,113,174,180,188,196,196
,202,206,210,213,220,227,243]

This vector represents pixels of the image region denoted in the matrix form by bold. To prepare vector representation for the boundary pixels (for example position (1,1)) we extend the image by replicating the neighboring n columns and n rows in vertical and horizontal fashion. An example of such replication for red channel is shown below. Now the element (1,1) of the previous red matrix of the intensity 48 has the required number of neighbors.

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

To avoid the outlier problem *p* extreme elements from both sides of each RGB vector can be rejected, making the clusterization problem more resistant to different type of noise, that might exist in the image. Assuming for example *p*=2 we get the previous vector in the form

$\mathbf{X}_2$=
[~~*30*,30~~,53,60,65,67,81,89,108,117,149,155,157,162,164,165,169,169,
173,181,197,213,215,~~222,240~~,
~~27,28~~,35,38,49,55,89,93,100,103,114,124,133,133,134,135,150,160,
178,189,195,220,238,~~249,254~~,
~~42,50~~,55,69,83,84,96,100,101,101,108,112,113,174,180,188,196,196
,202,206,210,213,220,~~227,243~~]

Searching for the nearest neighbors of the particular pixels we use the Euclidean distance between its vector representation. For example the distance between the marked RGB pixels of intensity (197, 49, 196) and the pixels of intensity (60, 93, 69) of the first row of matrix is equal d= 101.5185.

The alternative solution that will be compared in this work is application of the standard K-means algorithm, or K-mean algorithm preceded by the low-pass filtration of the image. The filtration might apply different filtering masks, for example averaging or Gaussian filters.

5. Results

The aim of numerical experiments was to make comparative analysis of the results of our automatic extraction of milk ducts to the classical K-means methods, all related to the results of medical expert. Fig. 4 presents three examples of original images containing the milk ducts of different shape and size.

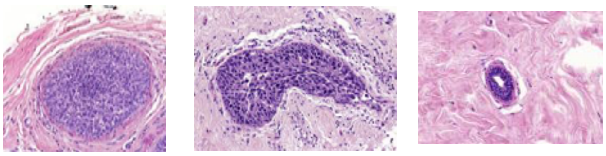


Fig. 4 The examples of original HE images of the milk duct used in numerical experiments

The results of our automatic system for these three images are presented in Fig. 5 (upper row). The boundary of the extracted images are clean and the background is without any artifacts. The bottom row presents the corresponding results obtained by using the standard K-means algorithm.

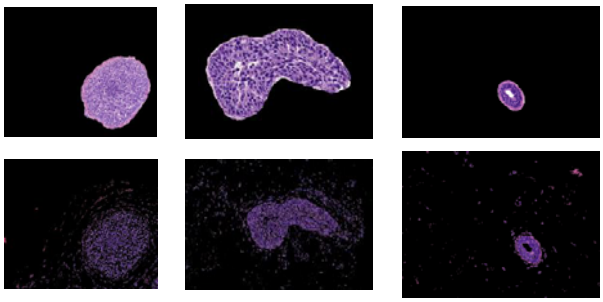
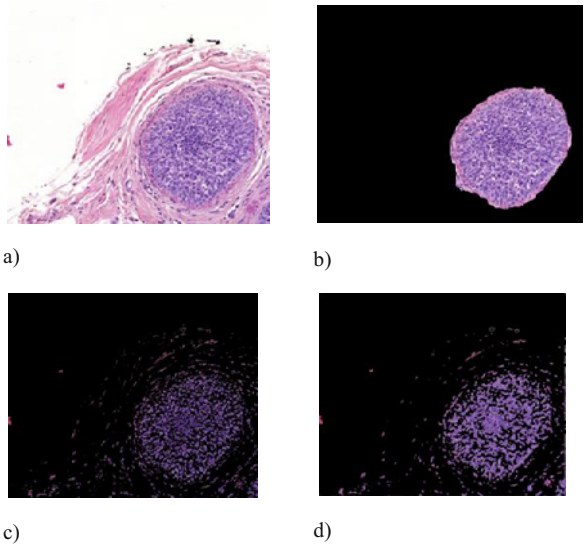


Fig. 5 The milk ducts extracted by our algorithm (upper row) and by standard K-mean (lower row)

The difference of the quality of extraction is evident. The results of our approach do not need any additional post-processing, while the application of standard K-means results in many artefacts, that should be eliminated by morphological operations.

Fig. 6 presents the details of graphical comparison of the extraction results of one chosen duct due to our algorithm and to the other existing approaches. Fig. 6a presents the original input image, Fig. 6b – the milk duct extracted using our method, Fig. 6c – the result of application of standard K-means and Fig. 6d – the standard K-means combined with application of the averaging filter with the 3x3 filtering mask.



a) b)

c) d)

Fig. 6 The images of the milk duct: a) the original input image, b) the milk duct extracted using our method, c) the result of application of standard K-means, d) the standard K-means with application of the averaging filter with the 3x3 filtering mask.

To compare the results in a numerical way the GT images prepared by the medical expert have been subtracted from the corresponding images created by the segmentation methods. The differential images have been defined on the basis of the binary masks. The sum of the nonzero pixels in the differential image was divided by the total sum of pixels in the GT image. The statistical results of such comparison have been done for the data base of 10 preparations, from which more than 80 milk ducts have been extracted. The statistical results are presented at application of different number of clusters used in K-means algorithm. Different variants of K-means (our approach, standard K-means and K-means with application of filtering) are depicted in Table 1.

Table 1 The relative difference of segmented image and GT image prepared by medical expert at application of different clustering methods.

Method	Number of clusters			
	3	4	5	6
Proposed method n=10	62,77%	**25,37%**	28,57%	39,82%
Proposed method n=5	67,14%	36,71%	**36%**	43,32%
Proposed method n=3	90,32%	39,32%	**39,18%**	42,17%
Average filter 3x3	141,83%	**41,37%**	58,61%	78,3%
Average filter 5x5	604,87%	514,21%	358,68%	**82,27%**
Gaussian filter 3x3	133,05%	**37,17%**	49,7%	75,08%
Gaussian filter 5x5	132,8%	**37,18%**	49,7%	75,95%
Standard kmeans	141,07%	**35,39%**	50,61%	66,35%

The values higher than 100% might have happened in the case when the size of the extracted ducts were much larger than the GT result.

The results show, that our method is superior to the standard K-means and to the K-means preceded by averaging or Gaussian filtration at different size of filtration mask. The advantage of the proposed method is very well seen in the best case, when at n=10 (mask size N=21×21).

6. Conclusions

The paper has presented the modified K-mean method for extracting the milk ducts in HE images of breast cancer. The main point of the method is substitution of the single pixel in K-mean algorithm by the region around this pixel and concatenation of RGB regions into common vector.

The proposed method has allowed reducing significantly the error of duct segmentation in comparison to the classical K-means approaches. Future research will be directed to test the method on larger population of HE images and compare the results with the GT patterns prepared by different experts. At the same time the numerical measures characterizing the basic parameters of extracted ducts will be proposed.

References

[1] C.Y. Chen, L.M. Sun, B.O. Anderson, Paget disease of the breast: changing patterns of incidence, clinical presentation, and treatment in the USA, *Cancer*, 107(7),1448, (2006).
[2] D.D. Dershaw, A. Abramson, D.W. Kinne, Ductal carcinoma in situ: mammographic findings and clinical implications, *Radiology*,170(2), 411–415, (1989).
[3] K. Anil Jain, Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, 31(8), 651–666, (2010).
[4] http://www.breastcancer.org/symptoms/types/dcis/diagnosis.
[5] C. Tse-Wei, C. Yi-Ling, C. Shao-Yi, Fast image segmentation based on K-means clustering with histograms in HSV color space, *2008 IEEE 10th Workshop on Multimedia Signal Processing* (MMSP), Cairns, Queensland.
[6] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to data mining*, Boston: Pearson Education Inc., 2006.