Computer-Aided Breast Cancer Diagnosis

Andrzej Obuchowicz

DPS 2013, Łagów Lub., Sept’11th, 2013

TEAM

NSC (NCN) grant: Breast Cancer Diagnosis based on Microscopic Images of the Fine Needle Biopsy Material

Institute of Control and Computation Engineering

University of Zielona Góra:

prof. Andrzej Obuchowicz, head

prof. Józef Korbicz

dr. Marek Kowal

dr. Andrzej Marciniak

Paweł Filipczuk

Department of Pathomorphology, Regional Hospital of Zielona Góra:

dr. Roman Monczak, spec. pathomorphology, oncological radiotherapy

dr Mariusz Kołodziński, spec. pathomorphology

Plan

- motivation
- breast cancer
- techniques of cancer diagnosis
- automatic diagnosis process
- image filtering and preparation
- pre-segmentation and segmentation
- sample results
- summary

Motivation

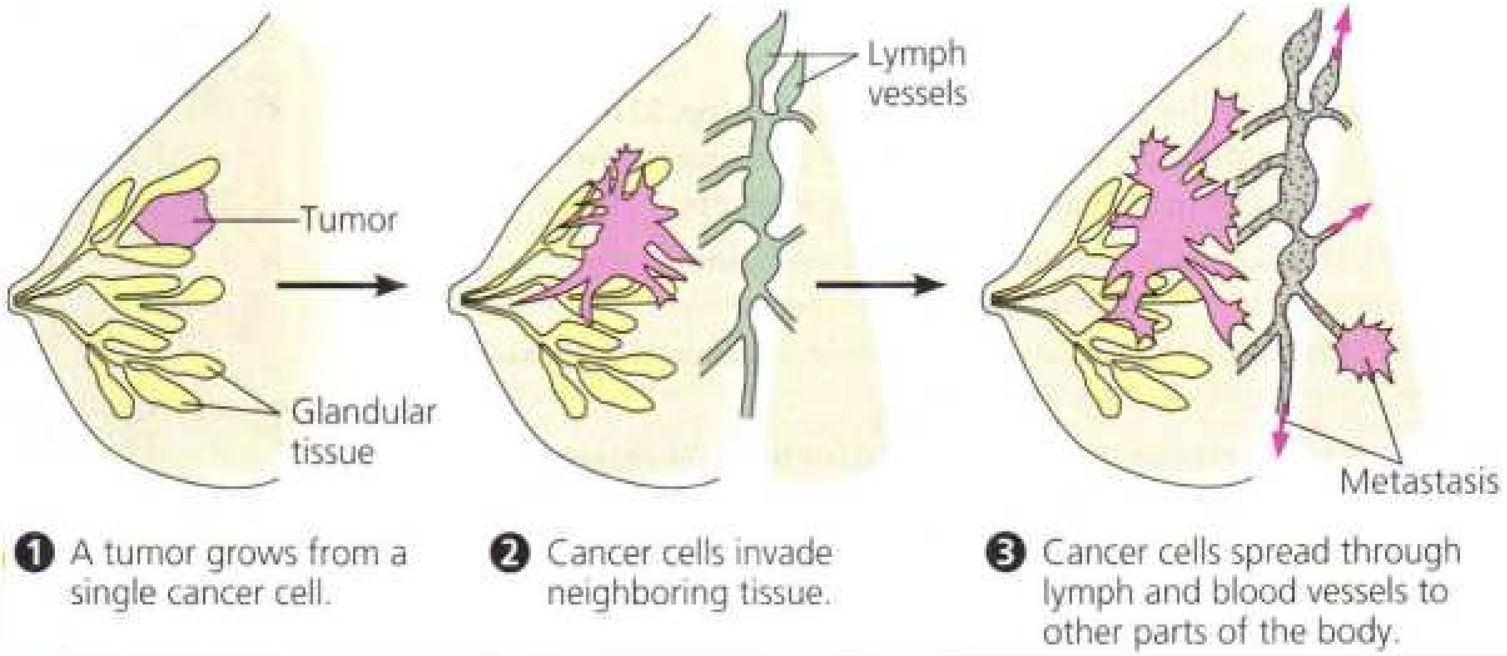
- **Breast cancer is the most often diagnosed cancer among middle aged women (about 1,400,000 cases per year).**
- **There are $7,600,000$ deaths due to cancer each year – $502,000$ caused by breast cancer.**
- **Poland 2008 – breast cancer: 14576 cases – 5362 deaths; the appearance frequency increases $3-4\%$ by year since 1980.**
- **Treatment effectiveness: depends on its fast detection in the early stadium.**

Motivation

- **Computer-aided diagnosis (CAD) – first approaches in mammography at the begining of 90's of the XX century.**
- **minimization of the human factor in diagnosis:**
  - **the specialist's mind can deform the essence of the massage included in the image by its own experience,**
  - **the diagnosis is independent on quality features,**
- **discovering new diagnosis rules, invisible to the naked eye:**
  - **decreasing of the fault diagnosis risk,**
  - **knowledge acquisition,**

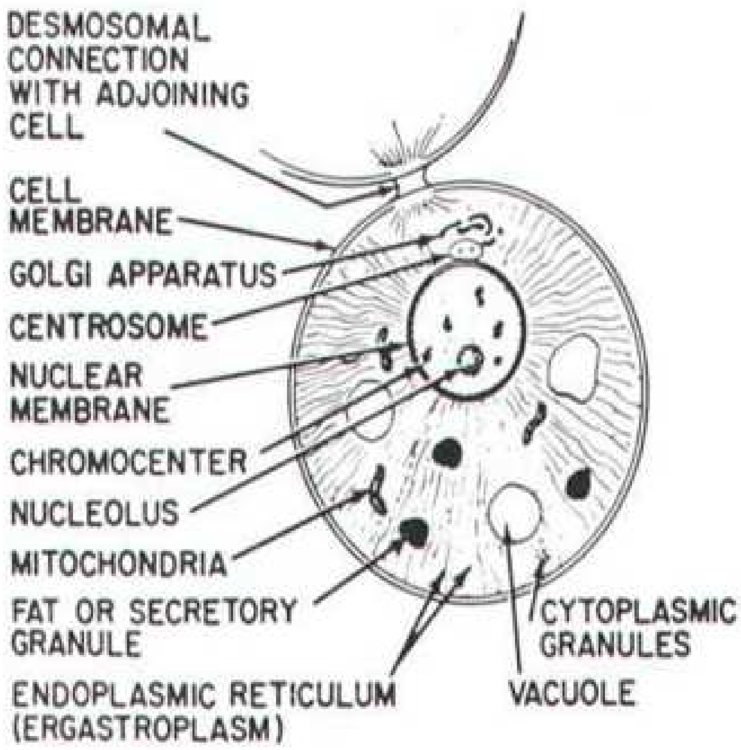
Breast cancer

- ✔ **Cancer cell is a normal cell in the tissue that underwent transformation.**
- ✔ **Tumor is a mass of abnormal cells within normal tissue. It is created when immune system tries to destroy cancer cell**

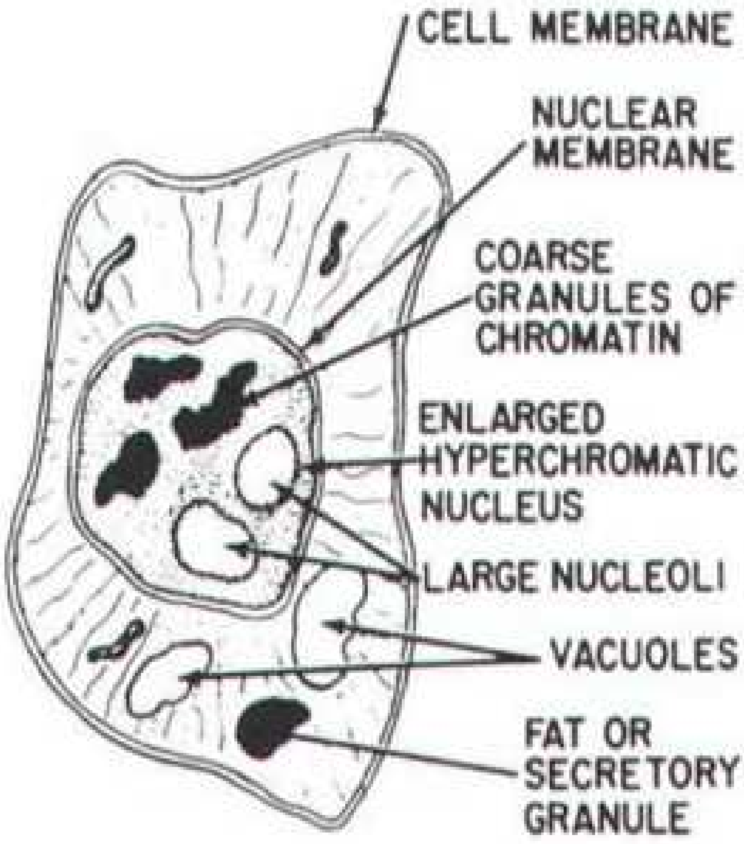


Healthy cell vs. Cancerous

healthy cell

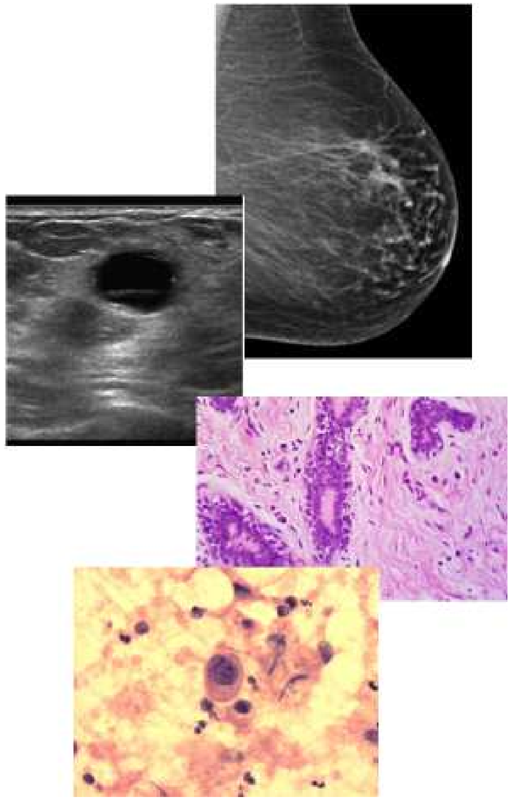


cancerous cell



Technics of cancer diagnosis based on the image analysis

- **imaging tests (x-ray mammography);**
- **histopathological diagnosis;**
- **cytopathological diagnosis.**



Radiological images

- **big area of a body**
- **detection instead diagnosis**
- **grey-scale images**
- **resolution (Computer Tomography)**
  $512 \times 512 \times 512$
- **over than** $134,000,000$ **voxels**

Cytological images

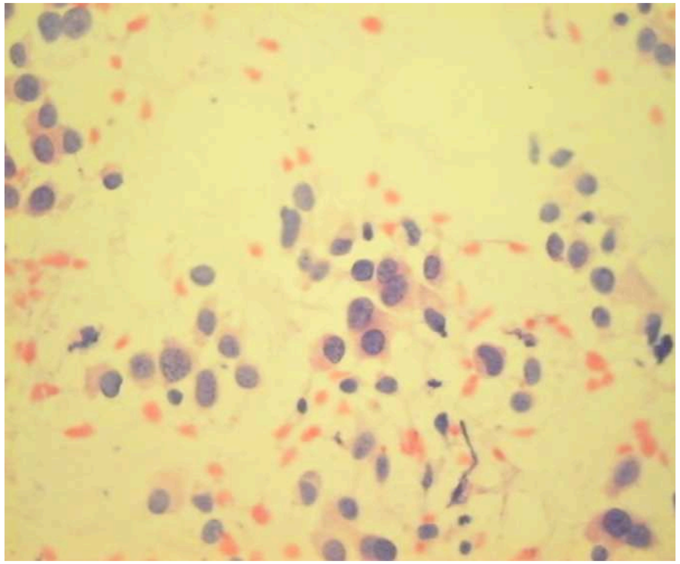
The diagnosis is based on the microscope image of cells taken from the body part with suspicious changes.

Techniques of the material acquisition:

- exfoliation technique,
- printing method,
- biopsy.

Characteristics of the material:

- disordered cells,
- well-isolated nuclei,
- resolution ca. $0.06\mu m$/pixel,
- ca. $50GB$ of information (virtual slides $\times 40$).



Histopathological images

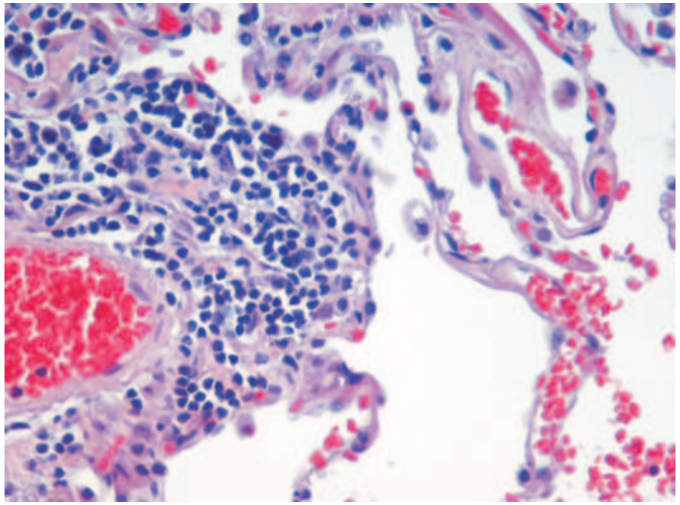
The diagnosis is based on the microscope image of tissue taken from the body part with suspicious changes.

Techniques of the material acquisition:

- **during the endoscopic test,**
- **during the surgical treatment.**

Characteristics of the material:

- **high precision of the diagnosis,**
- **visible tissue structure,**
- **visible lymphocytic tumor infiltration,**
- **resolution similar to cytological images.**



Triple test

triple test

- **physical examination**
- **mammography/USG**
- **fine needle biopsy (FNB)**

FNB images base no 1

equipment

- microscope: AXIOPHOT (Zeiss)
- camera: SONY CCD IRIS

Characteristics of the base

- number of cases: 75

  malignant: 25,

  benign: 25

  fibroadenoma : 25
- number of images: 750
- file format: RGB, 704 $\times$ 576
- focus: 10/160 $\times$ 2.5,
- number of colors: 16M

FNB images base no 2

equipment

- Olympus VS120 Virtual Microscopy System

Characteristics of the base

- number of cases (slides): 92
  malignant: 42,
  benign: 25
  fibroadenoma : 25
- file format: 200,000 $\times$ 100,000 (56 GB)
- focus: 40$\times$,

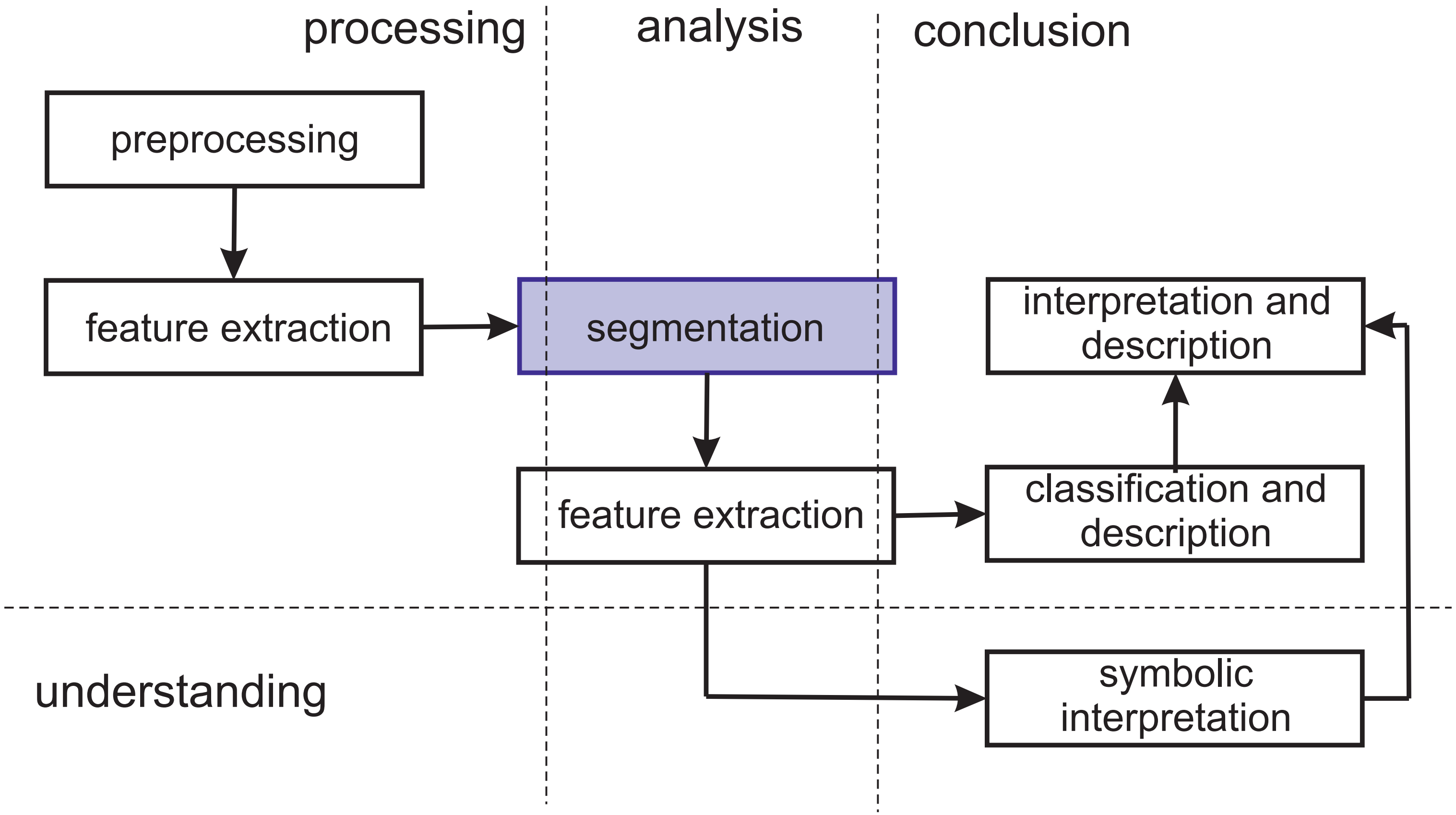
Additional base: 11 chosen areas for each case

- number of images: 1012
- file format: 1583 $\times$ 828

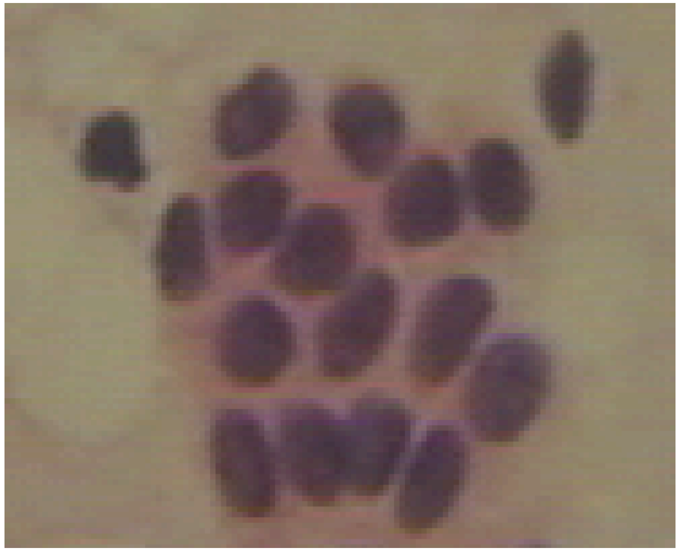
Problems of the automatic recognition

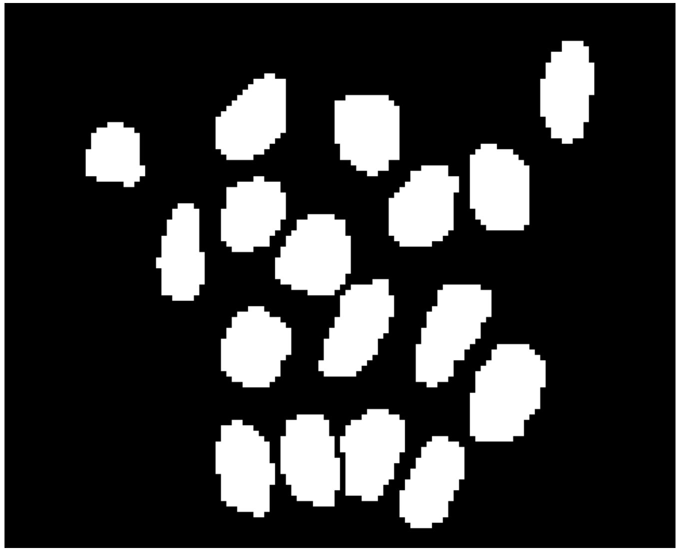
- **information uncertainty**:
  - imperfection of the acquisition process (noise, optical and chromatical distortion), resulting from the process of the material preparation, etc.,
  - nature of the image acquisition process (3D $\rightarrow$ 2D, scene lighting);
- **huge information to process** $\rightarrow$ space and time complexity,
- **necessity of the knowledge incorporation** (expertise, common sense knowledge).

Automatic diagnosis process



Segmentation





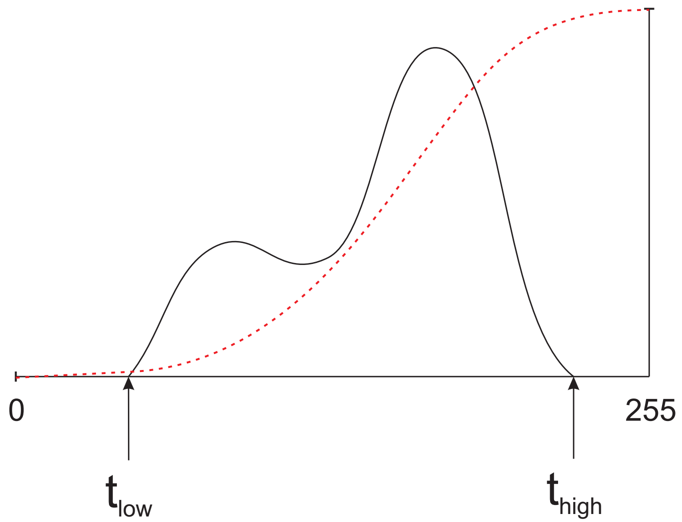
The basic problem – objective quality evaluation of segmentation

Preprocessing: contrast correction

- **Color elimination:** **luminance from the color space**

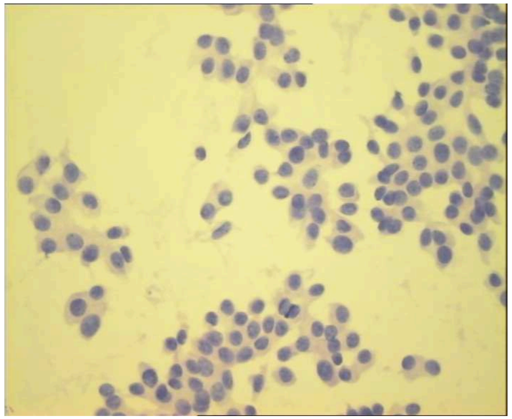
$$YCbCr\text{: } Y = 0.299R + 0.587G + 0.114B$$

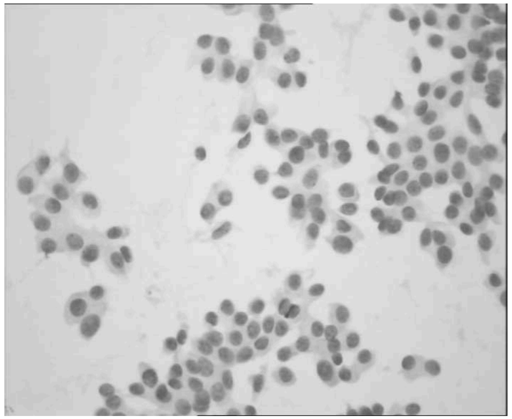
- **Contrast correction:**

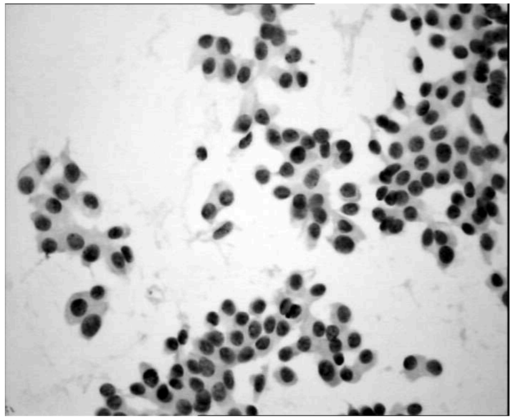


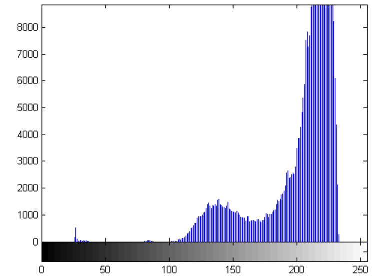
$$t_{low} = \arg_i^{(1)}\left(P(i) \geq \theta_{low}\right),$$
$$t_{high} = \arg_i^{(n)}\left(P(i) \leq \theta_{high}\right),$$
$$P(i) = \frac{\sum_{j=0}^{i} H(j)}{\sum_{j=0}^{N-1} H(j)}, \quad i = 0, \ldots, 255,$$

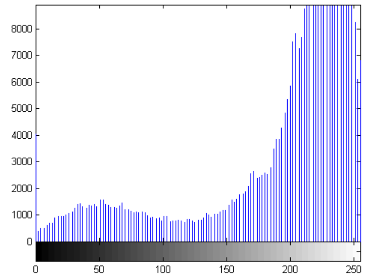
Preprocessing: contrast correction











Preprocessing: gradient masks

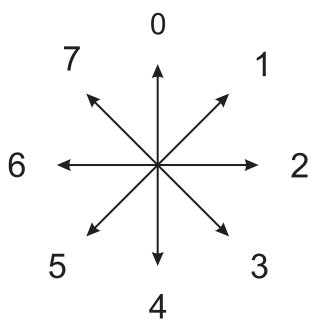
- **the gradient is estimated in eight directions in order to estimate circles, (basic masks: Prewitt's, Sobel's, ...),**

1	1	1
-1	-1	-1

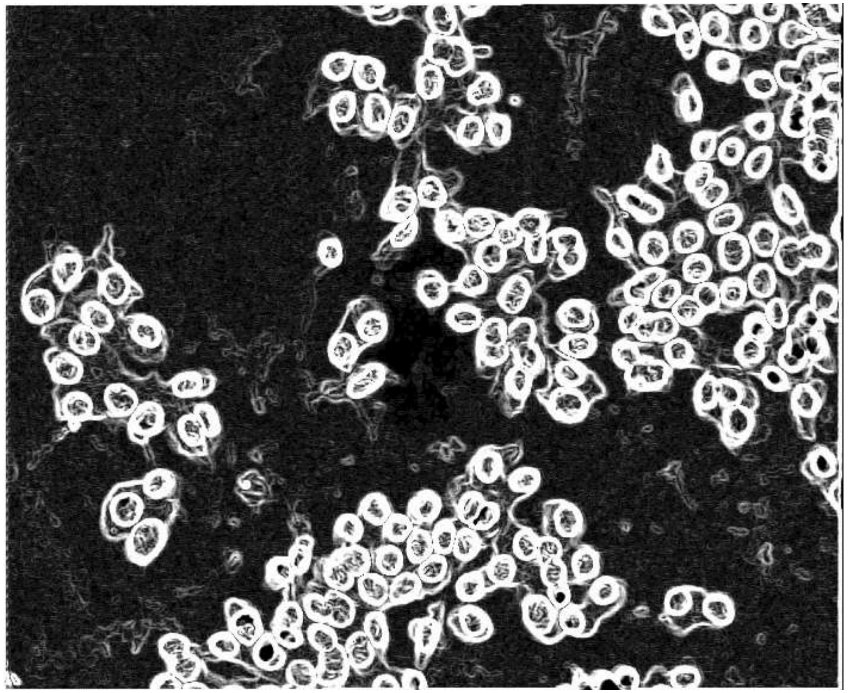
1	2	1
-1	-2	-1

3	2	1
2		-2
-1	-2	-3

1		
		-1

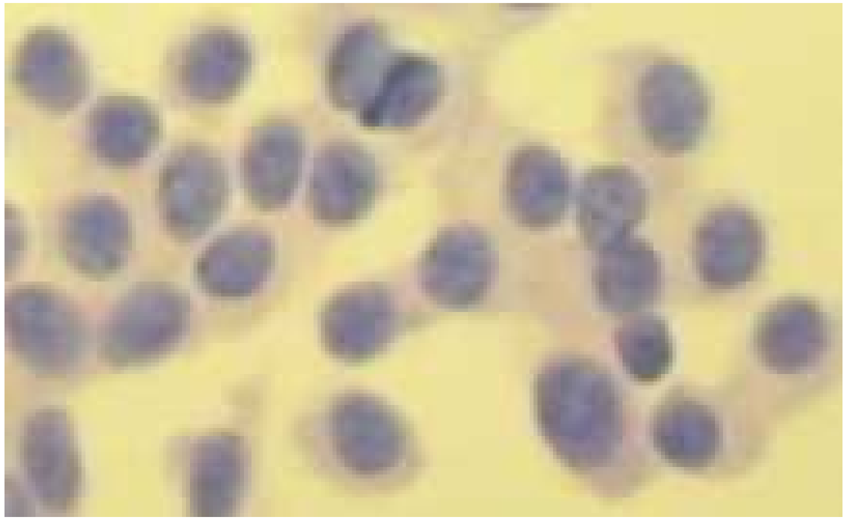


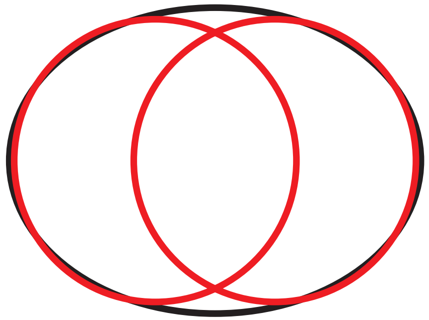
Sobel:



Preprocessing: presegmentation

$$\begin{cases} x = x_0 + a\cos(t) \\ y = y_0 + b\sin(t) \end{cases} \qquad \begin{cases} x = x_0 + R\cos(t) \\ y = y_0 + R\sin(t) \end{cases}$$





Preprocessing: Hough transform

- **circles detection with radius R, continuous form:**

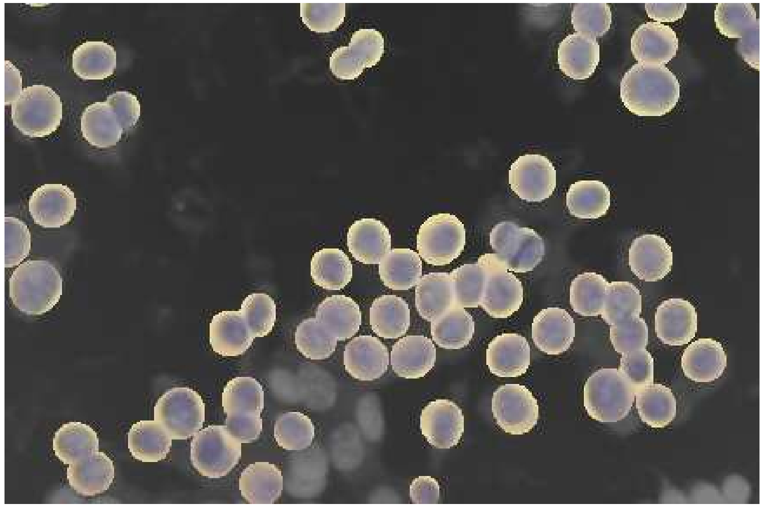
$$HT_{cont}(R,\hat{x},\hat{y}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\delta\Big((x-\hat{x})^2+(y-\hat{y})^2-R^2)\Big)dxdy,$$

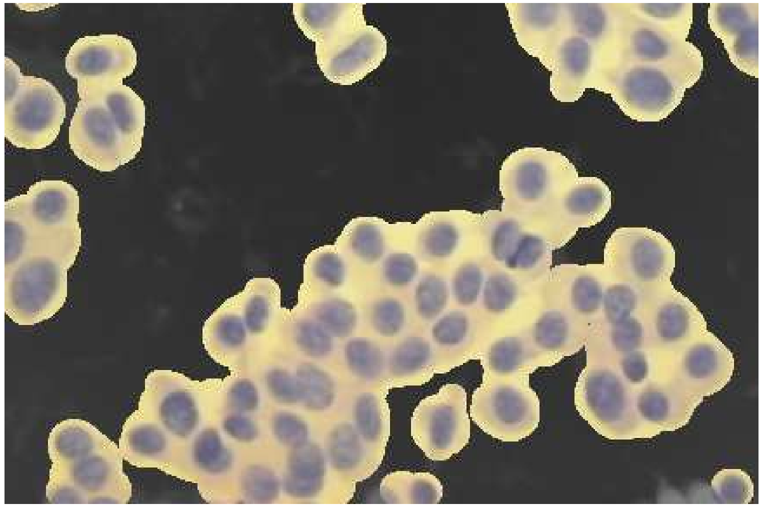
- **discrete form:**

$$HT_{discr}(R,\hat{i},\hat{j}) = \sum_{i=\hat{i}-R}^{\hat{i}+R}\sum_{j=\hat{j}-R}^{\hat{j}+R} g(i,j)\delta\Big((i-\hat{i})^2+(j-\hat{j})^2-R^2\Big),$$

Preprocessing: presegmentation example

⇛ HT **for** $R = 9$ **and** $R = 18$**:**



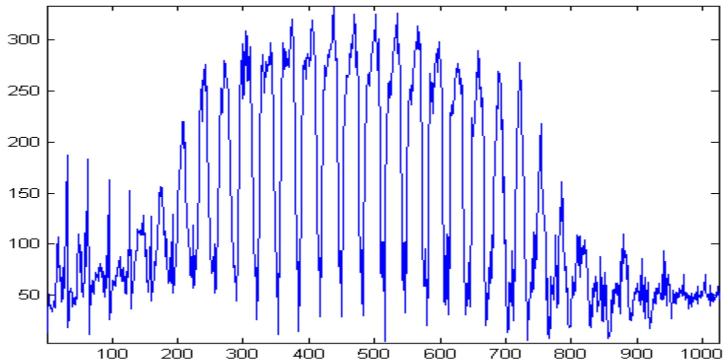


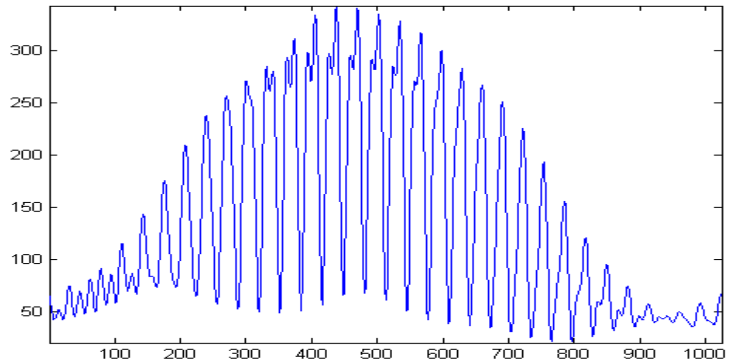
Preprocessing: nuclei shapes modeling

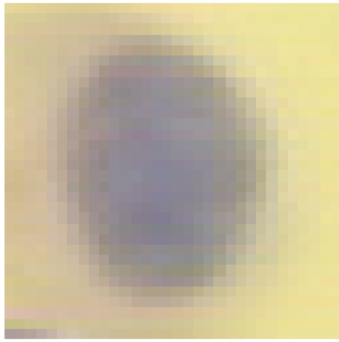
distance from the background:

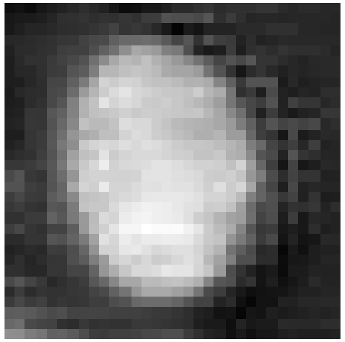
$$D_{euklid} = \sqrt{(I_R - B_R)^2 + (I_G - B_G)^2 + (I_B - B_B)^2},$$

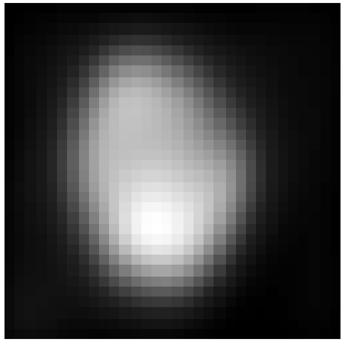
removal low frequences $F(m) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nm/M}$







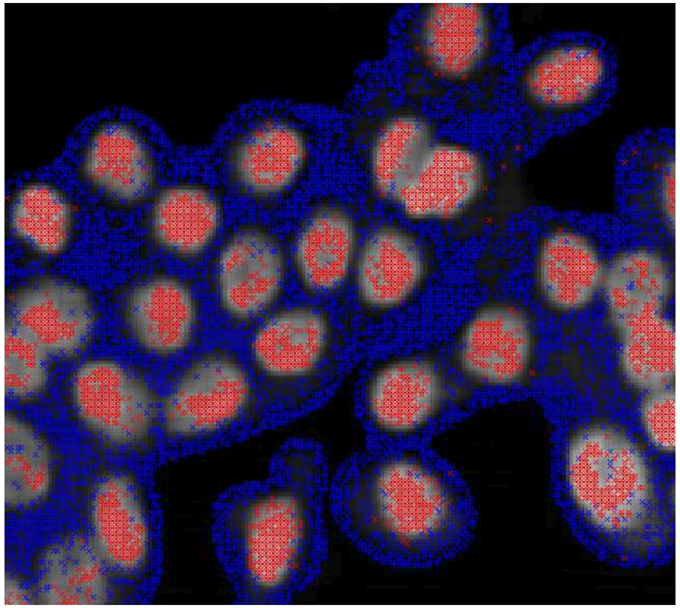


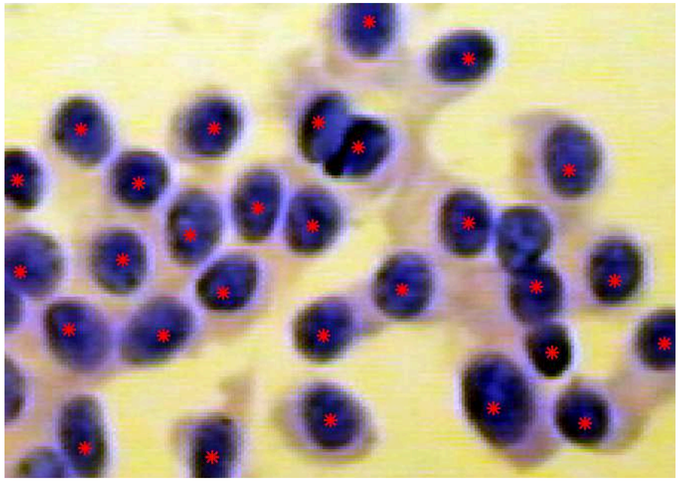


Preprocessing: nuclei localization with $(1+1)$ES

- **evolutionary strategy** $(1+1)$**ES**
  one-point version:
  - **two populations: first localizes background, second – nuclei,**
  - ϕ **– mean nuclei *high* (including neighborhood** $HT_{R_{min}}$**),**
  - **the nucleus is in the area, where population density is locally maximal,**

Preprocessing: nuclei localization with $(1+1)$ES





Nuclei localization using evolutionary strategies

Preprocessing: nuclei localization with firefly algorithm

idea

- ✔ technique inspired by the flashing behaviour of fireflies
- ✔ a map of luminance is treated as a map of an objective function: dark nuclei – valleys
- ✔ a swarm of simple agents which coordinate their local optima seaching taking into account their own fitness and attractiveness (brightness) of the other agents

Preprocessing: nuclei localization with firefly algorithm

- **the attractiveness of the agent (firefly) as a function of the distance r**

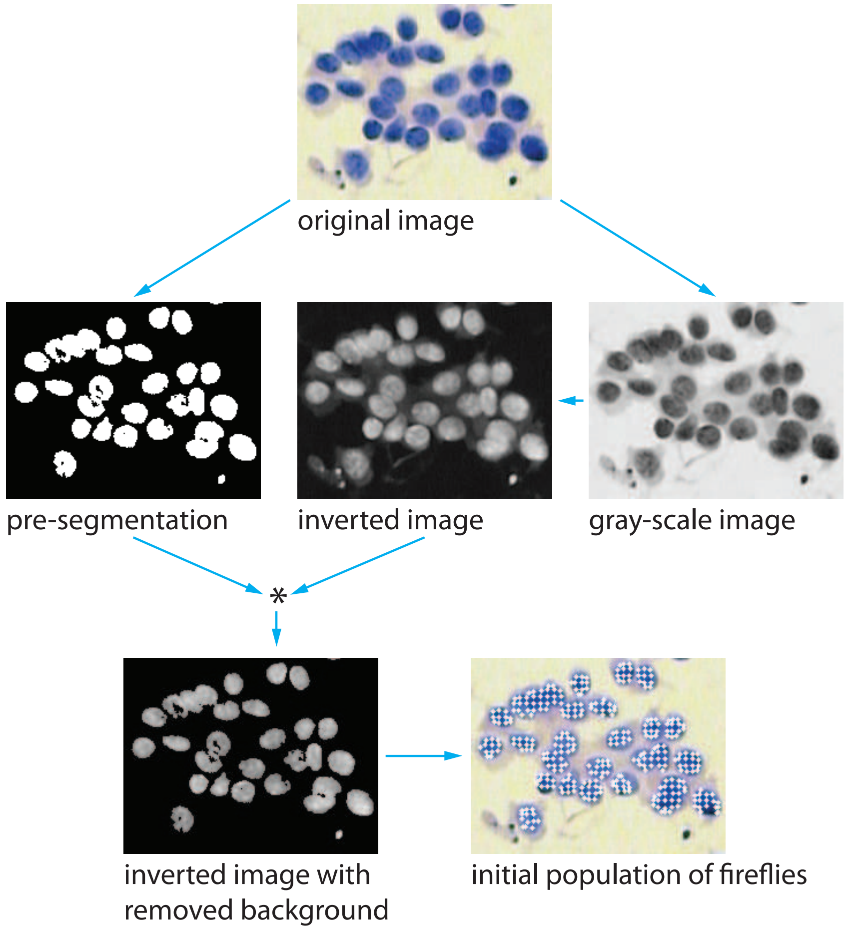
$$\beta(r) = \beta_0 e^{-\gamma r^2},$$

where γ - absorbtion coefficient; β_0 - attractiveness for $r = 0$ (depending on luminancy of the agent location and its neighbourhood)

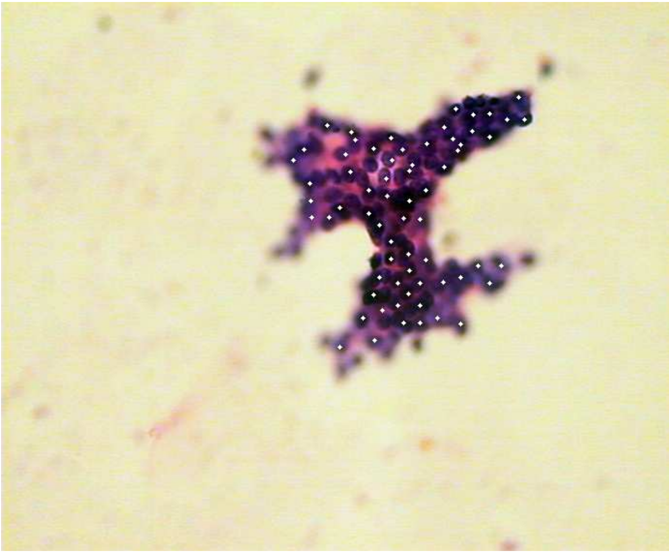
- **influence of the j-th firefly on the i-th firefly shift**

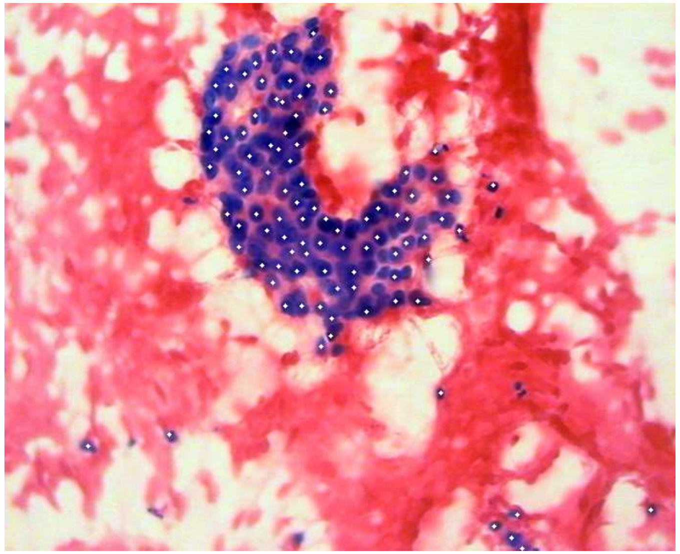
$$\boldsymbol{x}_i(t+1) = \boldsymbol{x}_i(t) + \beta_0 e^{-\gamma \|\boldsymbol{x}_j(t) - \boldsymbol{x}_i(t)\|^2} (\boldsymbol{x}_j(t) - \boldsymbol{x}_i(t)),$$

initialization procedure



Preprocessing: nuclei localization with firefly algorithm



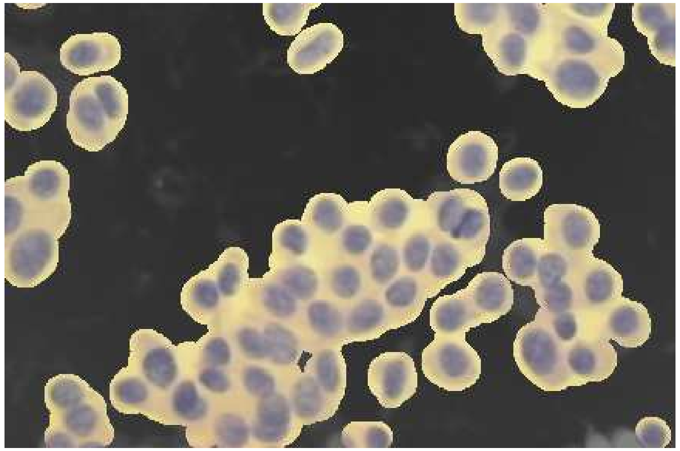
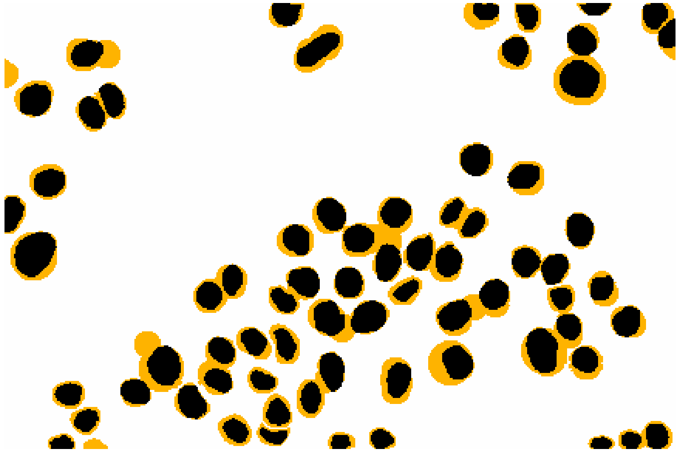


Segmentation: GrowCut

- **The algorithm is inspired by biological observations – growing and struggling to survive of bacteria population,**
- **bacteria spread from points of nuclei localization and edges of presegmentation masks,**
- **simple rule: each bacterium attacks its neighbors in each discrete step t,**
- **the attack strength is a function of the bacterium power θ_q and the distance between feature vectors $\vec{C}_q$ i $\vec{C}_p$ (the vector RGB of a given pixel),**
- **if the attack strength of the q-th bacterium is higher then the defence power of the p-th bacterium, then the label l_p of the p-th bacterium changes and $l_p = l_q$,**
- **stop criterion stable state of the bacteria system.**

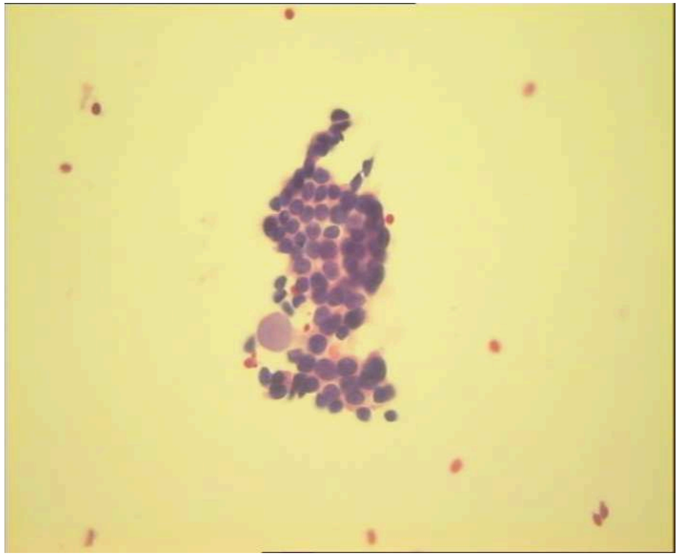
Segmentation: GrowCut

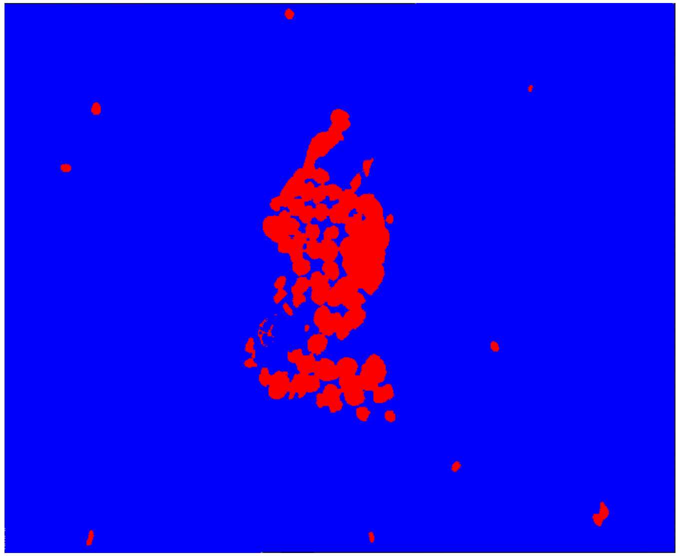
- **high quality segmentation for images with well-separated nuclei:**

Segmentation: GrowCut

- **problems and *mingling* in difficult cases:**





Segmentation – active contours

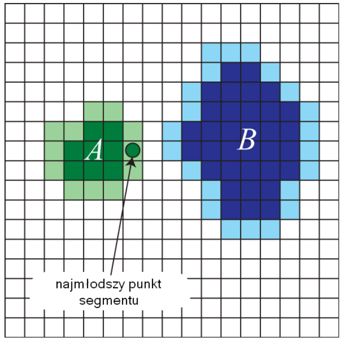
- **multilabel extension of the classical method,**
- **initiation: presegmentation mask and points of nuclei localization,**
- **growing speed of the contour is globally defined as:**

$$F = \frac{1}{|g(x,y) - \bar{g}(i)|^3 + 1},$$

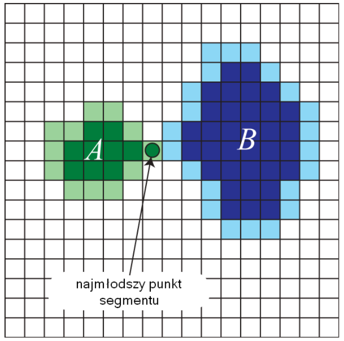
where $g(x, y)$ is a color of the contour, and $\bar{g}(i)$ is a mean color of the i-th segment – such a definition cases low speed of the contour near to the nucleus border,

Segmentation – active contours

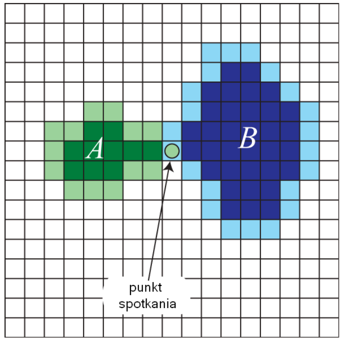
- **two segments, which meet,** **can join together, when the difference between their mean colors are under some threshold:**



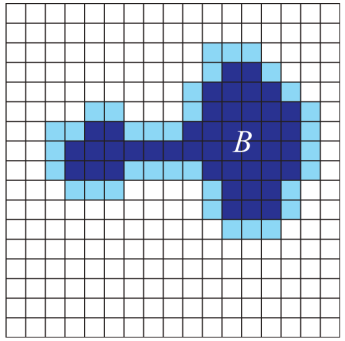
segments dismissed by one pixel



joined segments



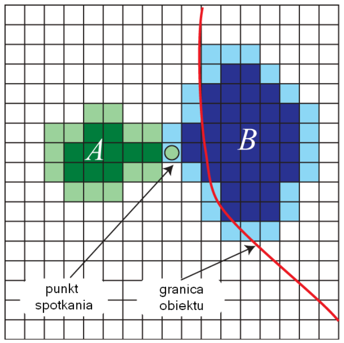
overlapped points



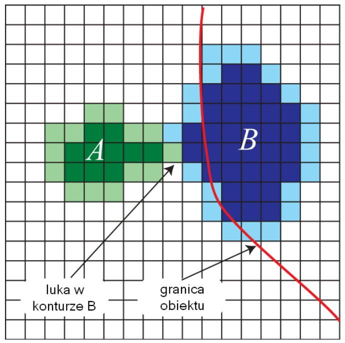
joined segments

Segmentation – active contours

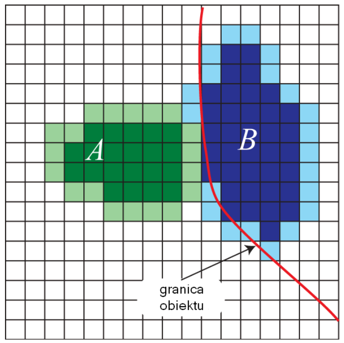
- **unjoined segment can push away the other one with lower difference between its color and the mean color of segments:**



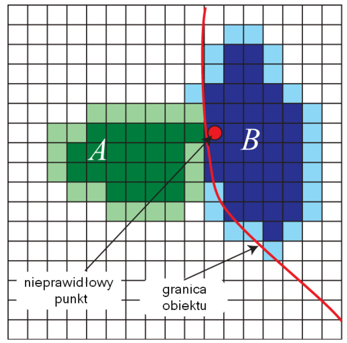
meeting point – differences calculation



the change of the contour point label



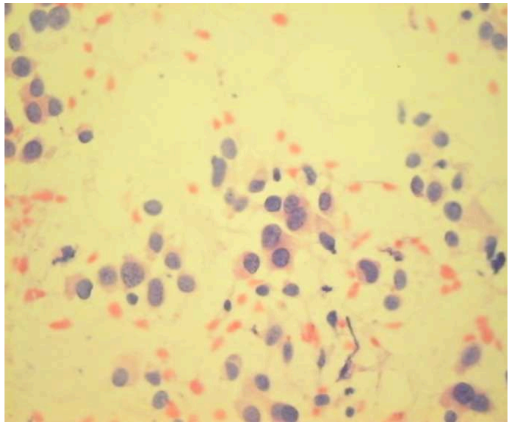
A on the nucleus border

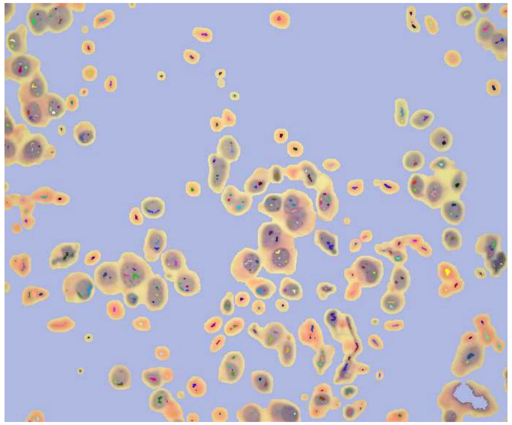


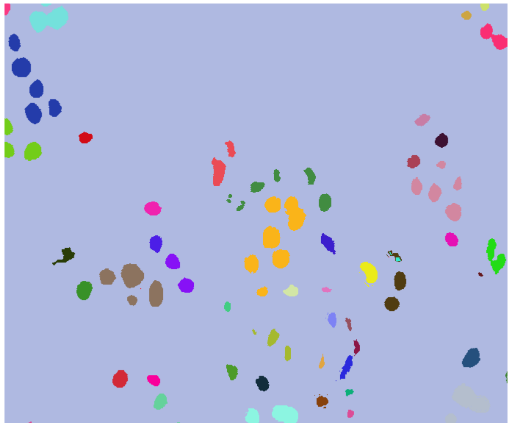
the point inaccessible for A

- **pushing away of the contour can take place only ones in order to avoid oscillations in this place.**

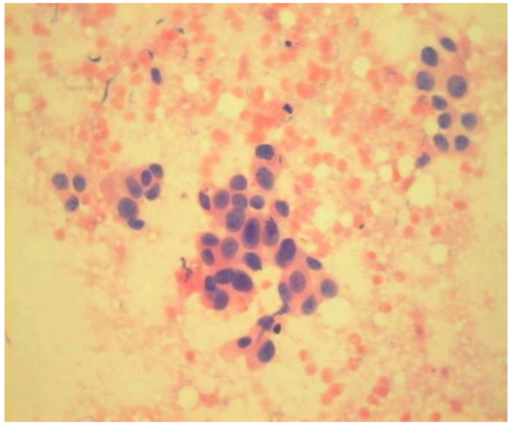
Segmentation – active contours

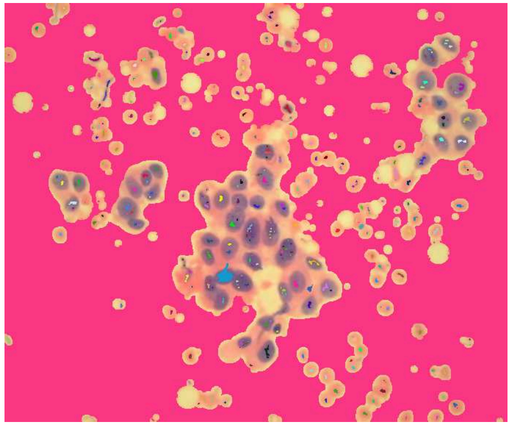


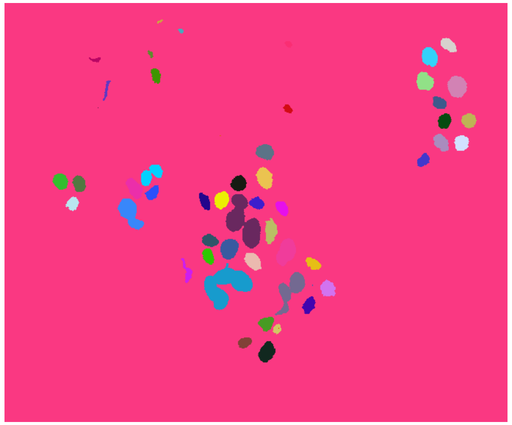




Segmentation – active contours

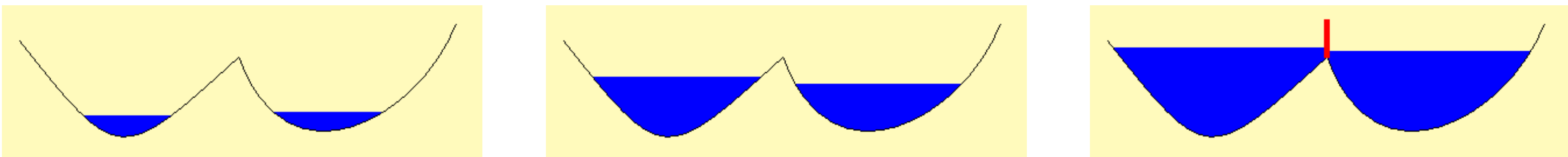






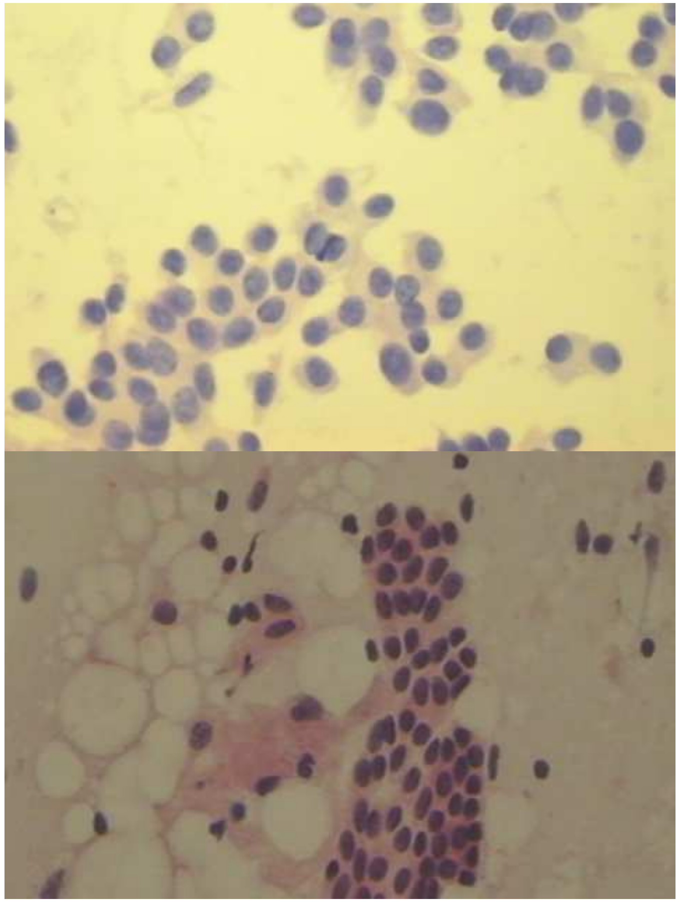
Segmentation – watershed

- **the algorithm inspired by nature (raining day in mountains),**
- **the image is treated as a landscape (nuclei are valleys),**
- **the landscape is flooded by the rain, ponds are created,**
- **dams separate meeting ponds,**

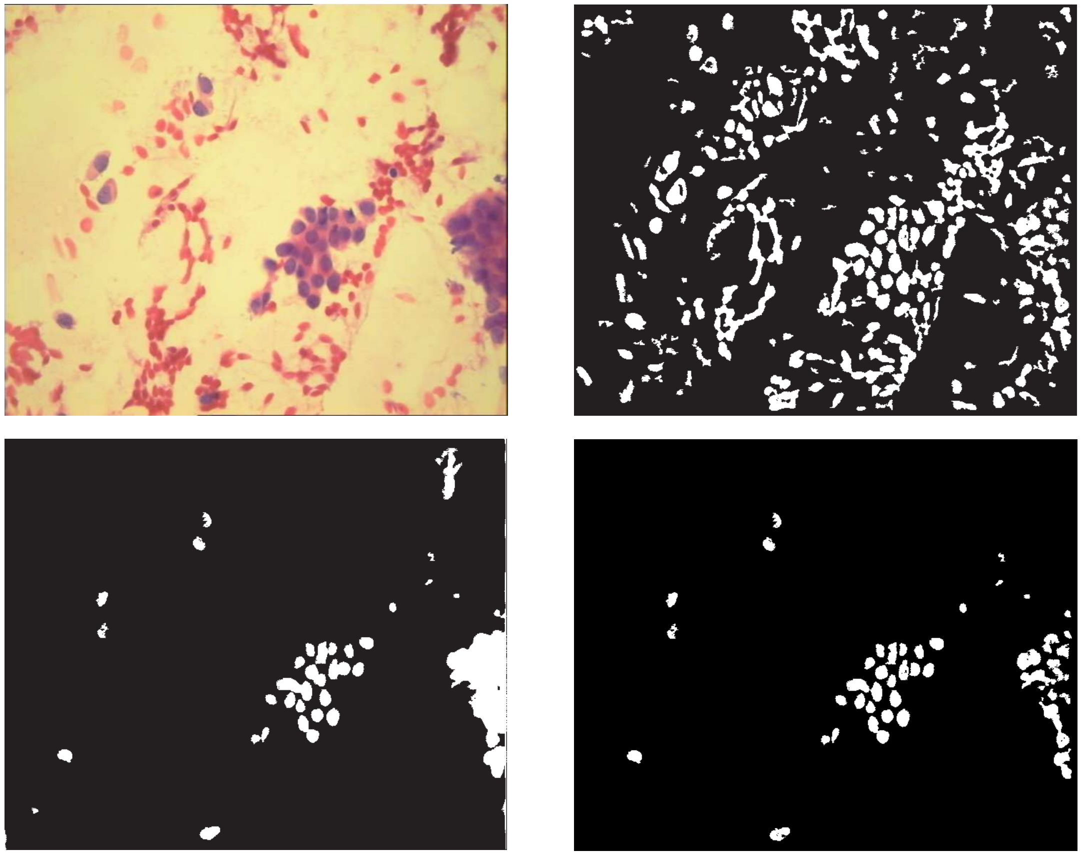


- **the nucleus is flooded to the meddle between *high* of the point of its localization and the mean *high* of the neighborhood background.**

Segmentation – watershed



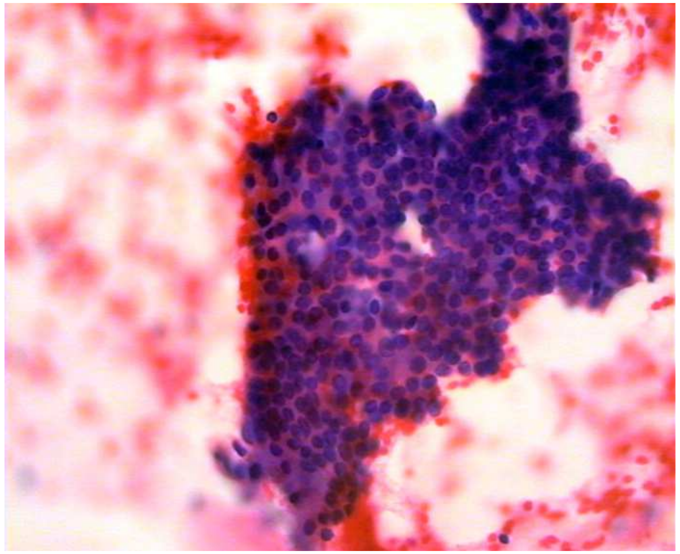
Segmentation – adaptive threshold + clusterization

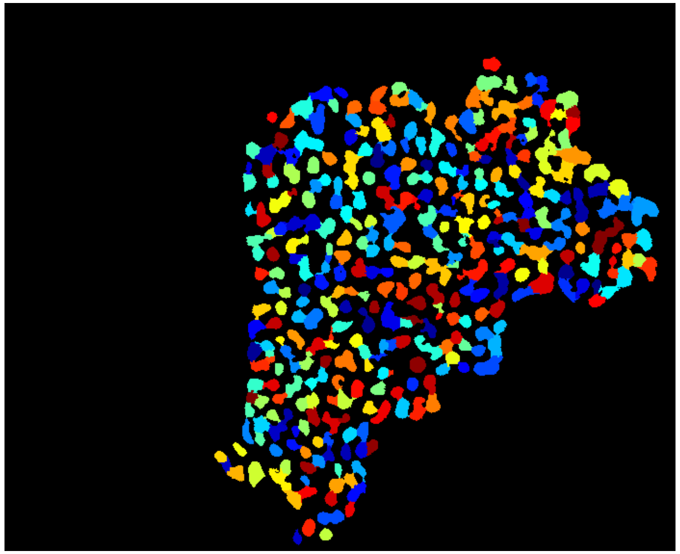


Segmentation – multilabel fast marching

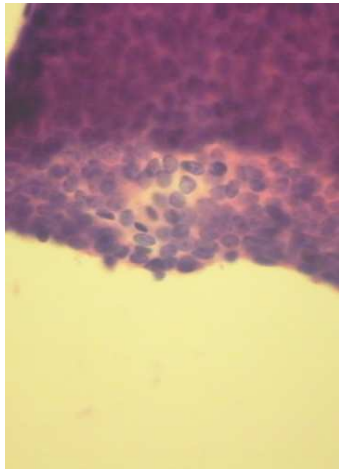
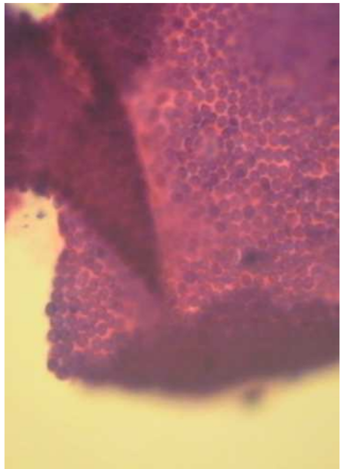
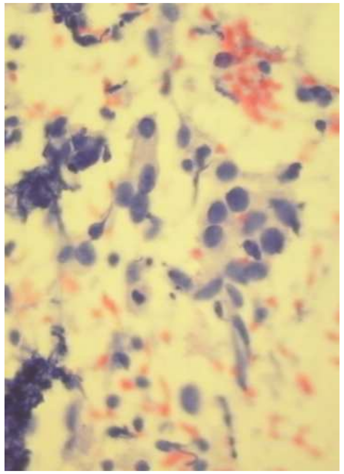
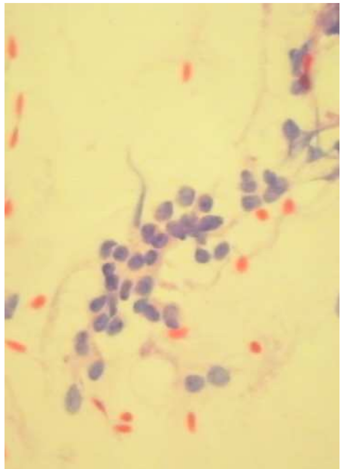
- **Fast marching method** is a special case of the level sets approach for monotonically advancing fronts.
- Algorithm starts with the initial fronts Γ_0^i. Next the fronts Γ^i evaluate with speed $F_i(x;y)$ in the normal direction where F_i are always either positive or negative.
- Front passes through a point $(x;y)$ at the time $T_i(x;y)$, that $|T_i(x;y)|F_i = 1$.
- Points $(x;y)$ with $T_i(x;y) < \theta_i$ (θ_i – threshold) describe the segmented object.

Segmentation – fast marching





Segmentation – difficult cases

Morphometric parameters

- morphometric parameters
  - area
  - circulation index
  - circumference
  - Malinowska's index
  - Blair-Bliss'es index
  - Danielsson's index
  - Haralick's index
  - Mz index
  - shape index Lp1
  - elliptical index
  - compactness
  - length of the shorter axis
  - length of the longer axis
  - eccentricity index
- statistical variables: mean, median, standard deviation, minimum, maximum ...

Example: CAD for FNB base 1

- **data basis:** 75 **pathological cases (**25 **malignant,** 25 **benign,** 25 **fibroadenoma)**
- **number of images:**9 **images for each case (summary** 675 **images)**
- **segmentation: adaptive threshold + K-means**
- **morphometric parameters discrimination analysis**
- **classification: kNN, naive Bayes, decision trees, ensemble classifier**

Result classification rate for selected sets of features

(m and v in brackets are mean and variance respectively)

Set of features	Classifier	Result
2 classes: benign, malignant		
area (m), distance to centroid (v), minor axis (m), minor axis (v)	kNN	100.0%
distance to centroid (m), eccentricity (v), perimeter (v)	naive Bayes	94.00%
distance to centroid (v), major axis (v), perimeter (m)	Ddcision trees	98.00%
minor axis (m), perimeter (m)	Eesemble classifier	98.00%

Result classification rate for selected sets of features

(m and v in brackets are mean and variance respectively)

Set of features	Classifier	Result
3 classes: benign, fibroadenoma, malignant		
area (m), area (v), distance to centroid (m), luminance mean (m), major axis (v), perimeter (v)	KNN	88.00%
distance to centroid (m), eccentricity (m), luminance mean (m), luminance variance (m), minor axis (m)	Naive Bayes	81.33%
area (m), luminance mean (m), luminance variance (m), major axis (m), minor axis (v), perimeter (m), perimeter (v)	Decision trees	85.33%
area (v), distance to centroid (m), luminance variance (m), minor axis (m), perimeter (m)	Ensemble classifier	88.00%

Result classification rate for selected sets of features

(m and v in brackets are mean and variance respectively)

Set of features	Classifier	Result
2 classes: benign plus fibroadenoma, malignant		
luminance mean (v), luminance variance (v), perimeter (v)	KNN	90.67%
luminance mean (m), minor axis (m), perimeter (v)	Naive Bayes	90.67%
area (m), distance to centroid (m), major axis (v), perimeter (v)	Decision trees	94.67%
area (m), area (v), perimeter (m)	Ensemble classifier	90.67%

Summary

- **CAD – human errors elimination, support in small hospitals (far away from experts), transformation: quality criterion $\rightarrow$ quantity criterion;**
- **medical images recognition are one of the most important element of the diagnosis process – necessity of searching more and more perfect solutions;**
- **artificial intelligence methods have a chance to explore unknown knowledge included in diagnosis images.**