Abstract: The detection of malignant cells in breast cancer needs a long and tedious screening. The Ad-Hoc (Automatic Detection of Healthy or Cancerous Cells) project aims at the automation of the diagnosis. Screenings assisted by computer are divided into two parts: the data extraction and its analysis. In this report, we are going to analyze the first part: the image segmentation.

The goal of this article is to show step by step the useful information contained in images, to demonstrate the difficulties and to propose a solution to segment the image.


L’objectif de ce rapport est de montrer pas à pas les informations que nous pouvons exploiter sur l’image, d’expliquer les nombreuses difficultés et finalement de proposer une chaîne de traitements permettant de segmenter l’image.

Keywords
Breast cancer, Cytology, Image segmentation
## Contents

0.1 Aim of the Ad-Hoc project ........................................... 3

1 Definitions ................................................................. 5
  1.1 Healthy cells .......................................................... 5
  1.2 Cancerous cells ....................................................... 6
  1.3 Spots : Where the images comes from ........................................... 6

2 The segmentation .......................................................... 8
  2.1 Background segmentation ............................................. 10
     2.1.1 Using Threshold and Opening ...................................... 10
     2.1.2 Results of the threshold and open-close filter .................. 15
     2.1.3 Using watershed .................................................. 16
     2.1.4 Results of the watershed based method .......................... 18
  2.2 Heap Extraction ..................................................... 19
     2.2.1 The heap problem ................................................ 20
     2.2.2 What to do with heaps ? ......................................... 20
     2.2.3 Isolated cells and heaps ......................................... 20
     2.2.4 Our algorithm for heap segmentation ............................. 21
     2.2.5 Results ............................................................ 22
  2.3 Extraction of nucleus positions ...................................... 22
     2.3.1 Last erosion ..................................................... 22
     2.3.2 Results ............................................................ 24
  2.4 Extraction of the nucleus boundaries ................................ 26
     2.4.1 Watershed ........................................................ 26
     2.4.2 Watershed using area closing .................................... 27
     2.4.3 Contour based method .......................................... 30

3 Further work ............................................................... 35
  3.1 Error Recovery ......................................................... 35
  3.2 Cell segmentation ..................................................... 35

4 Conclusion ................................................................. 37

5 Appendix ................................................................. 38
  5.1 Could we use another process ? ...................................... 38
     5.1.1 Our image processing ............................................ 38
     5.1.2 Other ways to segment the image ................................ 38
  5.2 Algorithm of the Ph. D of Lezoray (2000) ............................ 39
Introduction

A pathologist needs many years of studies to recognize a cancerous cell. The numerous criteria, configuration and the fuzzy chain of reasoning is hard as well for a human than for the computer.

This report won’t deal about reasoning, but only about a reliable way to extract data from the image for a future complex analysis. It will focus on the identification of three different objects: the background, the cytoplasm and the nuclei. Once the boundary of these objects is found, the analysis can be done. The goal of this report is to show our algorithm and to explain why it was chosen.

The two first chapters explain the context of our work, how it is related to the Ad-Hoc project, and some cytology vocabulary. Then our segmentation methods is explained: our work to extract the cells, the heaps and the nuclei is exposed.

0.1 Aim of the Ad-Hoc project

Breast cancer is a very common illness for women. Treatments exist, but the illness has to be found very soon. Every three years women should make a prevention test, and pathologists can propose a breast cancer test for any reason. Even if it is already high in France, the prevention programs should increase the demand in breast cancer detection in the next few years. So as to analyze the symptoms, a twenty minutes screening has to be done by a pathologist with a microscope. This screening is tedious, slow and subjective.

Our goal is to develop a fully automated algorithm for breast cancer detection. The automation of the process leads to many advantages:

- the pathologist can gain time on all the spots that are very cancerous or undoubtedly healthful;
- the computer gives objective results; thus the evolution of the illness can be better traced;
- the automation gives all the advantages of working with computers, such as an easy storage of the information, the capacity of sharing images between distant hospitals or the comparison between similar cases.

The automation is composed of two parts:

- The segmentation

This part leads to the separation of an image composed of pixels into different regions that correspond to a unique object. This is what we are going to explain in this article.

\(^1\)AD-HOC: Automatic Detection of Healthy or Cancerous Cells. The name of this project is subject to change.
• The decision

Once we have the size, the circularity and other data on the cell, an algorithm can lead to a diagnosis.

This part of the project may lead, for example, to the selection of the twenty more “cancerous looking” cells: the goal of the AD-HOC project is not to replace the pathologist but to help him.

This article won’t go into the decision part.
Chapter 1

Definitions

The goal of this chapter is to introduce a few words of specific vocabulary that are being used in this report.

1.1 Healthy cells

![A cell](image)

A breast cell is composed of a nucleus in a cytoplasm (Figure 1.1). Some informations seem obvious, but are important for the segmentation:

- most of the time objects are included in this order: \( \text{Concentration of chromatine} \subset \text{Nucleus} \subset \text{Cytoplasm} \subset \text{Background} \).

  Unfortunately, many contradictions can be seen such as background in the cytoplasm, nucleus outside the cytoplasm or concentration of chromatin outside the nucleus;

- the nucleus is darker than its cytoplasm; the contrast can be very low, and this is true only about the mean gray scale (the nuclei and the cytoplasm is not homogeneous);

- if the nuclei is very small (≈ 8 µ), then it can not be a cancerous cell. It might be a dead cell or a concentration of chromatin. Other information leads to such a conclusion. For
example if the cell is composed of one or more small nuclei, it might be a ploynuclear that has no relation with cancer.

There is no problem if the algorithm put such objects into the background partition. If the segmentation extracts them and put them in the nuclei partition, the analysis of the data should detect them.

Most healthy cell nuclei are small, the diameter is approximatively 10µ wide. These nuclei are also very close to a perfect circle, and the boundaries are smooth. Another important criterion is the cytoplasm size that is much bigger than the nucleus.

1.2 Cancerous cells

![Figure 1.2: Cancerous cells](image_url)

There are a huge number of small differences between a healthy cell and a cancerous cell. Diamond et al. (2001) use more than forty symptoms that could lead to the detection of cancer. Four criteria are more important than the others:

- cancerous nucleus diameter is huge (bigger than 13µm);
- the ratio nucleus/cytoplasm is much bigger than a healthy cell;
- the shape of the nuclei is irregular, which is obvious on Figure 1.2;
- the cancerous nucleus boundaries are irregular;
- cancerous cells are darker than healthy ones.

1.3 Spots: Where the images comes from

The images in this article come from spots. Spots creation begins with a fine needle aspiration in the patient’s breast. Of course the aspiration is done right inside the tissues that seem cancerous. This is the reason why, when a cancer occurs, most of the cells are cancerous, or at least they are numerous.

The second stage is the chemical destruction of objects that are not useful for cancer diagnosis, and that could create confusions. For example, the red corpuscles. The bath helps to
separate the cells from each other. A centrifuge is then used to plaster down the cells. The result is a spot, a 2 cm wide round projection of cells on a piece of glass. A microscope magnifies 400 times the spot, and a camera takes pictures (the resolution is 768x576 pixels).\footnote{Sometime I enlarged or shrink the images. Remember that a healthy cell is around 10\(\mu\) wide.}

Another method is the Papanicolaou method used on smears. This method has the advantages of having color, and is very popular. Our spots do not have any relevant color, but this new kind of extraction avoid the overlapping of the cells, and tends to be more and more used in France.
Chapter 2

The segmentation

The segmentation has to be done in three different levels:

• the image has to be divided into background and foreground (cells);
• the foreground is composed of cells. Each cell has to be isolated from another;
• each cell has to be divided into nucleus and cytoplasm.

We developed a new algorithm to segment images of cancerous cells (Figure 2.1), this algorithm will be divided into four steps:

1. The background segmentation.
   This step aims at creating a partition of the image into background and cells. This segmentation is done in two different ways: one based on watershed, and one based on a good threshold.

2. Heap extraction.
   The problem of the heaps is exposed, and this section explains the algorithm used to find heaps and isolated cells.

3. The research of the nucleus positions.
   The nucleus boundaries are hard to find, but are more easy to guess if the nucleus center is known. A method based on the last erosion is used to find the nucleus positions.

4. The research of the nucleus boundaries
   This is the trickiest part. Two methods are studied, one based on watershed, and one based on the boundary circularity.

At the end, further works is exposed, and after the conclusion a discussion explains why these steps were chosen and the possibility of having different steps. For example, it explains why nuclei were extracted before cells.
The segmentation

Figure 2.1: Our segmentation algorithm.
2.1 Background segmentation

The background does not interest us by itself. There is no information such as its color that can be used to establish a diagnosis. But once the foreground is known, the boundaries of all isolated cells are found. Moreover, if the foreground is an empty set, no treatments have to be done which can speed up the analysis.

Two methods have been used to separate the foreground and the background. The first algorithm that we created is based on thresholding. Once a good threshold has been found, some corrections are made using an open-close filter to find the background (Figure 2.2). The second method was described in Lezoray et al. (2000a). A subset of the cell region is found using a thresholding. Then a watershed is applied, using the subset as markers (Figure 2.2).

(a) Original, thresholding, open-close filter. (b) Original, thresholding and watershed result.

Figure 2.2: (a) First method. (b) Second method to find the background.

2.1.1 Using Threshold and Opening

The image is mainly composed of three things:

- nuclei (dark gray);
- the cytoplasm (gray);
- the background (light gray).

Figure 2.3: Top-left an isolated cell. Center a small heap. The goal is to extract them from the white background.
The first method is based on a threshold applied on the gray scale image. A threshold is a function that creates a binary image, in which the foreground is composed of the pixels higher (or lower) than a value.

In our images, it is possible to find a global value that is higher than most values of the cytoplasm and lower than most background values. After the threshold transformation, only a few pixels are wrongly classified. Only some pieces of heterogenous cytoplasm and some impurities in the background produce errors. We will discuss later how to get rid of them. The first difficulty is the automation of the process: an algorithm that works on images coming from any microscope has to determine the threshold value.

So has to find the grey scale that correspond to the background, the histogram can be analyzed (Figure 2.4). On the figure 2.4, a thin peak can be seen on the right-most part of the image. This peak correspond to the large amount of white color, and more precisely to the background. The local minima on the left of this peak can be extracted so as to determine the gray scale level that corresponds to the background.

![Figure 2.4: Histogram of Figure 2.3. The black peak on the right corresponds to the background.](image)

**K-Mean**

This algorithm is used for automatic classification. The user gives the number of partitions wanted, and the k-mean algorithm associates each element to a partition. The goal, of course, is to have homogeneous partitions. We have found various algorithms. This one is the basic one:

1. create $N$ clusters, and assign the elements to an arbitrary cluster;
2. repeat the following loop until the changes are small;
   - for each partition, compute the centroid of each cluster;
   - for each element $e$:
     - for each cluster $c$, compute $distance(e, centroid(c))$;
     - classifies $e$ to the closest cluster.

The distance function used between two gray scales is:

$$distance(g_1, g_2) = |g_1 - g_2|$$
2.1 Background segmentation

How many partitions should be asked?

One may think that the k-mean algorithm should be used with two partitions, due to the partition number needed (the background and the cytoplasm region). In fact, the gray scale level of the cytoplasm is very close to the background. The algorithm would make a partition containing both the background and the light part of the cytoplasm.

On Figure 2.5 a bump appears in the middle (gray). This bump corresponds to the gray scale of the cytoplasm. If the pixels corresponding to this bump are removed, two other bumps will appear (black and white line): There is three natural partitions. The gray one in the middle is wide (the cytoplasm is heterogeneous). The black line, that represents a small amount of pixels, corresponds to dark nuclei. The white line is thin (the background is homogeneous), but high. That is why the k-mean algorithm was used with three regions.

![Figure 2.5: Histogram of Figure 2.3](image)

As it can be seen on Figure 2.6, the k-mean does not create three partitions that correspond exactly to the background and both the cytoplasm and the nuclei. The cytoplasm is shrunk and nuclei are sometime erased (on the left), or extended over the cytoplasm (on the right).

![Figure 2.6: K-Mean algorithm run with three partitions on the Figure 2.3](image)

The increase of number of partition would make the result unpredictable. The k-mean algorithm might create two partitions for light and dark cytoplasm on an image, and creates two
partitions for the background on another.

Even if the result does not give us a good value of threshold, it can be used to calibrate the threshold.

**Open-close filter**

The two partitions done by the threshold do not correspond exactly to the background and the foreground. As you can see on Figure 2.7, two kinds of errors can be found:

- white impurities in cells lead to black\(^1\) dots on the cell area (it corresponds to the blue squares);
- the light does not shine with the same intensity on the spot\(^2\); pixels of the background are then in the cell partition.

![Figure 2.7: The threshold of the image leads to “salt” points (yellow) and “pepper” points (blue).](image)

To get rid of them an open-close filter is used, with a small circular structural element. The tabular 2.1 shows some morphological operators that can remove salt and pepper points. The open-close operator leads to the removal of both errors. The structural element is chosen very small (3 by 3), so as to avoid boundary alterations, and is circular. In the tabular, impurities have been enlarged. In reality, impurities are much smaller, and boundary deformations are sufficiently small for our work.

---

\(^1\)The black partition corresponds to the background.

\(^2\)Some black pixels on the background come from impurities
## 2.1 Background segmentation

### Table 2.1: Overview of a few mathematical morphology operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Output</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erosion</td>
<td>Any point belonging to the neighborhood has to be white</td>
<td>Suppresses all elements smaller than the structural element. Shrinks the boundaries</td>
</tr>
<tr>
<td>Dilatation</td>
<td>At least one point that belongs to the neighborhood has to be white</td>
<td>Suppresses all whole in the objects smaller than the structural element. Enlarges the boundaries</td>
</tr>
<tr>
<td>Opening</td>
<td>Erosion followed by a dilatation</td>
<td>Suppresses all elements smaller than the structural element. Boundaries are a little modified due to reconstruction</td>
</tr>
<tr>
<td>Closing</td>
<td>Dilatation followed by an erosion</td>
<td>Suppresses all holes smaller than the structural element. Boundaries are a little modified due to reconstruction</td>
</tr>
<tr>
<td>Open-Close</td>
<td>Opening followed by a Closing</td>
<td>Suppresses all elements and holes smaller than the structural element. Boundaries are a little modified due to reconstruction</td>
</tr>
</tbody>
</table>

Table 2.1: Overview of a few mathematical morphology operators
2.1.2 Results of the threshold and open-close filter

The result on Figure 2.8 corresponds to the partition needed. The heap boundaries correspond exactly to the one found, the dark pixels on the background have been removed. The two holes than can be seen in the cell partition are not created by impurities, but correspond to the background of the original image.
2.1 Background segmentation

2.1.3 Using watershed

This method based on watershed is an adaptation of the one describe in Lezoray et al. (2000a). The watershed algorithm has been proven very efficient for segmentation in Vincent and Soille (1991). This method is based on the variation of the gray scale. Thus, it depends less on the fact that all the cytoplasm is darker than all the background.

First we are going to see what happens when the watershed is used without gradient, and then the Beucher gradient will be used.

Watershed without gradient

The image has to be inverted, so has to make the water comes from the background (which is more homogeneous than the cells). The problem is that the water won’t necessary be stopped by the cytoplasm if the gray scale changes little by little (Figure 2.9).

![Watershed without gradient](image)

Figure 2.9: Watershed without gradient. The water is not necessary stopped at the beginning of the cytoplasm (left).

Watershed with gradient

The watershed algorithm creates boundaries on high peaks that delimit regions. We want to detect the cytoplasm boundaries. Indeed, we use a gradient that creates peaks (white on the image) on the boundaries. The Beucher gradient was used to find the boundaries, where the gray scale is changing. The Beucher gradient is the arithmetic difference between the dilatation and the erosion (Figure 2.10):

Erosion for a gray scale image \( f \) with a structural element \( B \):

\[
\epsilon_B(f)(x) = \min_{b \in B} \{ f(x + b) - B_v(b) \}
\]

dilatation:

\[
\delta_B(f)(x) = \max_{b \in B} \{ f(x + b) + B_v(b) \}
\]

Beucher gradient:

\[
\rho_B = \delta_B - \epsilon_B
\]

---

3You can see the differences in the appendix.

4The white background corresponds to high values. Inverted, it correspond to low values.
Figure 2.10: The Beucher gradient is used for edge detection.
2.1.4 Results of the watershed based method

Figure 2.11: Result of the background extraction using watershed.

As you can see it on Figure 2.11, the results correspond to the segmentation needed. But
some problems remains, such as on Figure 2.12. The problem is due to the Beucher gradient. Figure 2.13 explains what happens: the transition between the cytoplasm and the background is soft, and boundaries are more in the inside of the cytoplasm.

Figure 2.12: The light and heterogeneous cytoplasm fools the watershed. The red crosses correspond to errors

Figure 2.13: From the bottom to the top: gray scaled image (black), gradient(red); blue: the water being stopped too late.

2.2 Heap Extraction

Heaps are concentration of cells(Figure 2.14). Heap sizes differ from one to another (sometime hundred time bigger than in this figure; it can fill all the image). Inside heaps, the cell detection is hard or impossible.
2.2 Heap Extraction

2.2.1 The heap problem

The cell concentration makes them hard to separate: boundaries between two cells are fuzzy and the contrast is very low. Most of the time, heaps are composed of several layers. The camera focus makes some cells fuzzy, and cell superpositions lower the contrast. Most information inside a heap differ from the ones in isolated cells, for example the average gray scale is higher.

2.2.2 What to do with heaps?

Heaps are hard to segment, and various strategies are possible:

- The first one is to ignore them. This strategy relies on three facts:
  - Most of the time it is impossible to detect the cell boundaries inside a heap, or even to count the cells. Heap analysis is error prone.
  - Heaps have a lot of differences from isolated cells. Not only the concentration differs, but the gray scale, the number of dead cells, the size of the cytoplasm, and most criteria. That lead to many problems in the analysis of the data extracted from the isolated cells and the heaps. It might be better to have reliable measures (from isolated cell) well understood, than having more data, but less trusted.
  - In breast cancer diagnosis, the pathologist is looking for a recurrent symptoms. It is very unlikely to have cancerous cells in the heaps, and only healthy isolated cells.

- Even if the segmentation inside heaps is hard, the extraction of some features such as the average gray scale are easy. With an automated classification of the gray scale, it is even possible to have a quite good idea of the ratio of the cytoplasm and nuclei area. Even if the extraction of the data is poor, some informations can lead to show malignant heaps to a pathologist.

- Pathologists do not ignore the heaps nor try to view all their details. Even if they can’t see all boundaries, they will try to find the cells that are the more contrasted. Then, information are extracted for these cells only. Even if the extraction is hard and error prone, it can makes the little difference between two diagnosis.

2.2.3 Isolated cells and heaps

To extract all connected components bigger than a value, the opening is often used. But the reconstruction will not re-create the isolated cells sticked on the heap (Figure 2.15). The difference between wide connected components and their reconstructions would lead to many small
errors. On the top left of Figure 2.15, a nucleus has been cut by the opening, and its cytoplasm has been added to the heap partition. Such errors are not inconsiderable, because the cytoplasm area is a main criterion for breast cancer detection.

At this stage of the segmentation, cell boundaries are unknown. Thus, our algorithm keeps isolated cells stucked on the heap in the heap partition.

Figure 2.15: If isolated cells are not considered (such as on the left), some cytoplasm pieces are cut or added (red cross). We would like to add these isolated cells (right) to the heaps.

2.2.4 Our algorithm for heap segmentation

Our algorithm to find the heaps is:

1. connected components are searched; the input binary image comes from the output of the background / foreground segmentation;

2. an erosion is done on the same image. The structural element is very wide, at least bigger than the bigger isolated cell;

3. All connected regions that
   * still exist after the erosion belongs to the heap partition.
   * are destroyed by the erosion belongs to the isolated cell partition.

   Note that “isolated cells” does not mean that cells do not touch any neighbor. It means that the neighborhood is small and do not correspond to a high concentration of cells.

The operator is an opening and not an area opening used in the section 3.4.2. This was chosen because we are looking for wide heaps. If an area opening would be done, some isolated cells linked by thine objects (such as a DNA strand unrolled) could have been set in a wrong partition.
2.2.5 Results

The partition is correct as it can be seen on the Figure 2.16. Moreover one can easily choose the minimum size of the heaps. That can be useful if we need to segment the image at different scales. Isolated cells correspond to all cells that do not belong to the heap partition.

2.3 Extraction of nucleus positions

Nucleus boundaries are hard to find. This is why we divided this step into two parts. The first one is the seek of nuclei; we propose an algorithm to have the nucleus positions. The positions will then be used to find the nucleus boundaries in the second part.

2.3.1 Last erosion

To find the nucleus position, the last erosion is used. This method has been exposed in Lezoray et al. (2000a).
If you erode an object thinner than the structural element, it is removed. The last erosion operator finds for each connected component the last subsets removed by the erosion (in practice it corresponds to the local maxima of the distance map).

![Image erosion](image.png)

**Figure 2.17: Image erosion.**

The erosion is not done on the foreground partition, with a threshold on the original image. The threshold image should be a subset of the partition created by nuclei. Of course, this subset should have at least a pixel of each nucleus. The fact that it is a subset makes the threshold easier to determine than the previous one.

The last erosion can not be used directly on threshold images: the heterogeneous cytoplasm has dark regions that creates small dark points. If the last erosion is applied on these dark points, it will create a false cell position. To avoid such an error, several erosions are going to be done on the image at first.
2.3.2 Results

The first important thing is that there is at least one position for each cell. The second one is that nucleus positions are often in the center or at least not on the edges. The result corresponds to the nucleus position, even if some errors can be observed.
Why is there some errors?

The fact that there is sometime more than one marker per cell is due to three reasons:

- Heterogenous nuclei create irregular connected components when the image is thresholded. That trigger, sometime, to more than one cell position.
  
  At this segmentation stage it is really hard to correct it (Figure 2.19). Even if markers are very close to each other, it can be an odd nucleus configuration (for example, a heterogeneous nucleus, and a chromatin out of its nucleus).

  ![Double nuclei.](image)
  ![Double nuclei?](image)
  ![Single nucleus. But the gray scale is very close to a double nuclei.](image)

  Figure 2.19: The double nuclei detection creates false positive and false negative cases.

- Heterogeneous shapes of the nuclei(Figure 2.34). Cancerous cell shapes are not necessary circular. That fools the last erosion which creates many markers.

- The third reason is the cell sizes. Wide cells are often heterogeneous, and the erosion often cut wide cells into different parts (Figure 2.20).

  ![a) Wide cancerous cells. The algorithm found many positions.](image)
  ![b) Small cells. (same scale as a).](image)

  Figure 2.20: The algorithm detects small and healthy cells. Wide heterogeneous cancerous cells are harder to segment.

Some positions are found by our algorithm outside the nuclei. This is due to the low contrast. One of the criteria to be a cell is that its nucleus is darker than its cytoplasm, itself darker than the background. As you can see in Figure 2.21, the cytoplasm is not darker that all nuclei (it is true locally). Moreover, the hierarchy (background, then cytoplasm, and then the nucleus) is not respected: nuclei can touch the border of the cytoplasm (or even out of the cytoplasm), or the cytoplasm can be empty.

Sometime, the threshold sets some dark object in the nucleus subset. The low contrast makes it very hard to avoid (Figure 2.21).
2.4 Extraction of the nucleus boundaries

In this section we will explain how to find a nucleus boundary for each position found. Two methods were used to extract them, one based on watershed, and one based on the boundaries.

2.4.1 Watershed

This method is based on watershed transformation. It is an adaptation of Lezoray et al. (2000a). As you can see in the appendix, Lezoray uses the color to disambiguate situations in which the cytoplasm looks like a nucleus. Our images are mainly composed of purple and white colors; even if the three components are different, no significant color space has been found. Therefore we worked more on the image simplification. The goal is to enhance all important pixels, such as pixels inside nuclei. At the same time, we have to take care of the thin boundary and to spare them.

At first, the segmentation was done in three steps:

1. the contrast was enhanced;
The segmentation

Most of the time, boundaries between the cytoplasm and the background are less contrasted than boundaries between the nucleus and the cytoplasm. But this rule has some exceptions. Therefore, we enhanced the contrast on the image to increase the contrast of the boundaries corresponding to the dark objects and to decrease the contrast of the light objects. The function used was:

\[ f(x) = \text{MAX} - (\text{MAX} - x)^2 / \text{MAX} \]

where \( \text{MAX} = 255 \).

2. the Beucher gradient was applied on the image to have a boundary map;

3. Watershed was used; the markers were the nucleus positions.

Homogeneous cells are correctly segmented using this method. But as you can see in Figure 2.23, this method does not work well if cells are heterogeneous.

![Figure 2.23: Zoom on errors created by our method without area closing. The area closing resolves the problem for the two heterogeneous cells on the left.](image)

Markers are likely to be found in darker regions of cells. For the cell on the top left, markers were found in the two concentration of chromatin. The water of the watershed is likely to be stopped by the gray scale changes between the chromatin and the nucleus. This method confuses the nucleus boundaries and the chromatin boundaries. Hence we need to remove all dark impurities inside nuclei. The area closing was choosen to perform this operation.

2.4.2 Watershed using area closing

So as to have more homogeneus region, the area closing was used. The impurities are now removed before the watershed transformation. To remove these dark impurities on gray regions, the closing procedure comes in mind. But a regular closing blurs and enlarges the boundaries (Figure 2.24). The size of the structural element could be changed, but

- if it is too small, it does not clear impurities at all;
- if it is too big, boundaries are destroyed.

Thus we used an area closing, that removes all connected components with an areas smaller than a value, independently of their shapes. The nucleus boundaries are unchanged, because nuclei are wide. This algorithm is described for gray scale image in Vincent (1992).

\footnote{I call impurities all dark objects that make cells heterogeneous. For example the concentration of chromatin produces dark dots.}
2.4 Extraction of the nucleus boundaries

Figure 2.24: Area closing creates homogeneous surfaces and does not damage the boundaries.

Results

Figure 2.25: Watershed with area closing results

The result is over segmented, due to a number of marker too high. 52 nucleus boundaries are correct, 7 nucleus are segmented but the boundaries are not accurate. These false boundaries are not circular at all, and the data analysis is likely to detect these errors. 5 pieces of cytoplasm are set in the nucleus partition. They are all smaller than a nucleus, thus their detection by the analysis should be possible.
In theory, one cancerous cell is enough to determine the illness of a woman and the result here shows that most boundaries are accurate. Knowing that if a cancer occurs, cancerous cells are numerous, the results are reliable.

The few errors appearing in the segmentation can be explained:

- False nuclei found correspond to small dark pieces of cytoplasm. These errors are not going to be recovered by the segmentation, but their area could be used to detect them.

- The focus is not correct on all images (some cells can be closer to the microscope). Moreover, dead cells and other objects blur the spot. The shadow on the nuclei makes them hard to extract. Sometimes it is even impossible to extract their size (Figure 2.26 [a]). It is easy to know if cells are fuzzy of not with the measure of the gradient through the boundaries. This kind of information can be added to other data on the cell for further analysis.

- The over segmentation is due to the too high number of markers. One cannot easily remove these markers because of the double nuclei. On Figure 2.19 the common points of the double nuclei and heterogeneous nuclei are showed. Many heterogeneous cells are over segmented.

- The low contrast causes errors. Very wide impurities (such as on the figure 2.21 [b]), can lead to an error. Watershed may be stopped by the high gradient around impurities, and our method might set the nucleus in the cytoplasm partition, while setting the impurity in the nucleus partition. This kind of error can be detected during the data classification: impurities are smaller than nuclei (10 µ). If the cell is cancerous, we will lose a clue for breast cancer detection. But cancerous cells are numerous or not present, thus the detection can be done on other cells.

- Cell textures are very heterogeneous. It leads to many high gradient levels inside the cell, which produce over segmentation and sometime segmentation errors. The over-segmentation will be discussed in the further work, section 3.1.

Figure 2.26: Fuzzy cells. Fuzzy boundaries create error in the segmentation (b and c). Most fuzzy cells are dead (a).
2.4 Extraction of the nucleus boundaries

A heterogeneous nucleus in a heterogeneous cytoplasm.  
A heterogeneous nucleus with an odd shape.  
The overlapped cells produces heterogeneous regions.

Figure 2.27: Textures create error during the research of the nucleus position and during the watershed segmentation.

**Drawbacks of this method**

The problem with this method is that it does not use the shape information. It leads to some errors that look obvious for a human. Therefore, we tried another kind of segmentation based on the shape information.

Figure 2.28: These boundaries are found by our algorithm based on watershed. They do not correspond to a nucleus shape.

### 2.4.3 Contour based method

The radius methods consists in drawing radiuses around the cell center, and to search the path that goes through the boundaries (it means, were the gradient is high), and that is circular. This method is presented in Bamford and Lovell (1999), and was originally used to check if the segmentation of healthy cells is correct.
The segmentation

Figure 2.29: The algorithm tries to maximize the path cost, in term of circularity and boundary matching. The green path will be preferred to the red one, due to its circularity.

Three parameters are used for this method:

- $\alpha$ coefficient: this coefficient determines the importance that the user give to circularity. If this coefficient is too high, only healthy cells are well segmented, and if it is too low, shapes likes on Figure 2.28 are accepted. The cost of the path is:

  \[ \text{Cost} = \alpha \times \text{circular\_shape} + (1 - \alpha) \times \text{boundary\_matching} \]

- $\Delta$angle: on Figure 2.29 $\Delta\text{ang} = \pi/4$. This determines the path accuracy. In Bamford and Lovell (1999) 16 radiuses are needed, but we are going to show that we need a lot more radiuses.

- range: boundary are not search on the entire image. The range gives informations on the nucleus: the smallest size possible, and the biggest size possible. In Bamford and Lovell (1999) 12 nodes per radiuses are used, but we need a lot more.

Because we are working on cancerous cells, the nucleus sizes differ a lot (Figure 2.20). Heterogeneous textures (Figure 2.27 lead to error in the seek of nucleus positions. As shown in Figure 2.30, range and radius numbers have to be high.

Figure 2.30: The nucleus position is not centered (due to the heterogeneous texture). The cancerous cells are not circular, and are huge.

In Bamford and Lovell (1999), all possible paths are computed (all paths going clockwise, it makes $12 \times 16 = 192$ paths). But in our case, the number of different paths is much bigger,
2.4 Extraction of the nucleus boundaries

and each path can not be computed. In fact, we work with at least 64 radiuses, and a range of [5..54] (at least), that makes a minimum of 3200 different paths per cell. Therefore, it is too slow to compute circularities and costs of all paths. Moreover some paths highly not circular cost more than a circular path out of any boundary. Right now, we use a Dijkstra path finding to find the best path, by introducing a cost if the path is not locally circular. This algorithm is going to be changed soon by another algorithm based on snakes.

**Edge enhancement: Nagao Filter**

The main problem is that our fuzzy images do not look like the image on Figure 2.30. We needed to enhance edges and make regions homogeneous. Therefore, we used a Nagao filter, which smoothes the image while preserving edges. More precisely, the Nagao filter works with a 5 by 5 neighborhood, with 9 windows (Figure 2.31). For each input pixel, the output pixel is equal to the mean of the more homogeneous window (the one that has the smallest variance).

![Nagao filter windows](image)

**Figure 2.31: The Nagao filter uses 9 windows. One centered, the four corners and the four sides.**

Around nucleus edges, the amount of color decrease from the center to the outside. Near the edge on the inside, the Nagao filter will not choose the gray scale existing on the edge (changes in the gray scale make it heterogeneous), but it will choose a window inside the nucleus (Figure 2.32). Boundaries are then more contrasted.

![Original image vs Nagao filter](image)

**Figure 2.32: The Nagao filter creates homogeneous regions while preserving edges.**

To sum up this method of segmentation:

- first the cell positions are found, using last erosion;
- then, the Nagao filter is applied on the image;
for each position, the best boundaries are computed. The circularity and the boundary matching are used to determine the boundaries. In fact, the matching of the boundaries includes two things:

- the gray scale should be lower outside the boundary. This implies that the Beucher gradient is not needed;
- the gray scale should be dark inside the nucleus. The function that enhance the contrast of the nucleus boundaries is not needed.

That speed up treatments, because the gradient and the contrast enhancement is just computed where it is needed. Moreover, the derivative is more accurate than the Beucher gradient, because it is compute in the direction of the radius. The Nagao filter could be computed only for the pixels needed too.

Results

Figure 2.33: In black radius intersection and boundaries found by the radius method.

This method produces less nucleus boundary errors, as it can be seen on Figure 2.33. This contour detection seems to correspond to the need of our project.
All the nucleus extracted have the features of a nucleus. It can turn to a drawback, because errors should be more difficult to find by the data analysis than the obvious errors of the watershed based method. Another drawback is the number of parameters. It has more parameters than the first method. Moreover, if the snake algorithm is used this number will increase, with the initialization and the energy parameters.

What makes the boundary extraction better, is the information of circularity. But some cancerous cells are not circular at all. If you look at the cell on the upper left Figure 2.34, you will see that the algorithm find a circular nucleus, even if the nucleus is not circular. But if a nucleus is very cancerous, other nuclei should be cancerous but with a more circular shape.

Figure 2.34: The circularity energy is hard to manage.
Chapter 3

Further work

3.1 Error Recovery

This section aims at working with over-segmentation in the watershed method (Figure 3.2. This might be done in two steps, as it is explained in Lezoray et al. (2000a): merging and splitting.

Results of the watershed method can not be used directly: the over segmentation has to be recovered. To avoid the over segmentation, the regions that share a boundary can be merged. That could correct the error 1 of Figure 3.2.

By merging different region, the natural (and correct) nucleus boundaries between two nuclei that share a boundary is destroyed (the merging would destroy the boundary of the number 3, Figure 3.2). Because it is hard to choose if the region should or should not be split, the problem won’t be avoided; but we could correct it. Once merged, the regions could be split using a distance map.

A watershed could be used on the distance map (such as Figure 3.1). Using local maxima as markers, watershed would create different regions corresponding to one nucleus.

![Figure 3.1: Distance map: Further it is from the sides, the more it is green. The red crosses correspond to the local maxima (and to the markers of the watershed. The watershed splits the double nucleus into two nucleus (black line)](image)

3.2 Cell segmentation

Our method to segment the background, heaps and nuclei allows us to extract the majority of the features. But the boundaries of the cells in a neighborhood are not detected. Thus, one
criteria can not be compute now, it is the ratio $\text{sizeof}(\text{nucleus})/\text{sizeof}(\text{cytoplasm})$. In Lezoray et al. (2000a), the research of cell boundaries is done in four steps:

1. nuclei are extracted; they are going to be the makers;
2. an exponential filter is applied on the image, to enhance the boundaries between the cells;
3. a Beucher gradient is used on the filtered image;
4. the watershed transformation is applied on the image, using the makers.

We tried this method but it was not efficient on our images. This segmentation seems possible on some parts of the image, but the low contrast makes the boundary detection difficult on some parts of the images(Figure 3.3). Changing in this method should be found so as to complete the segmentation.
Chapter 4

Conclusion

If some errors can be seen on the segmentation, they are rare and understandable. Most errors are obvious, and can be rejected by the analysis of the lack of circularity, and by their odd sizes. Moreover in case of breast cancer, cancerous cells are numerous, and the overall data of our segmentation can be trusted.

In most articles, cell images come from Papanicolaou smears. Therefore, colors can be used to increase the difference between the object, which lead to better boundaries. Even if our image have sharper edges, we do not have such information, and our algorithm creates partitions that correspond to our needs. Generally, the segmentation described is used for healthy cell segmentation, which all have the same circularity, homogeneity and size. Our algorithm extracts all the boundary of these healthy nuclei.

In fact, the segmentation of cancerous cells is different to the segmentation of healthy cells. The methods have to give to the segmentation enough freedom to accept strange cells, and to avoid at the same time any objects to belongs to a wrong partition.

Of course, the segmentation might be a less good than the one done by a pathologist that spend 7 years to learn medical science. But the goal of the Ad-Hoc project is to help the pathologist, not to replace a human. Using this segmentation one can easily determine the wider nuclei, locate the darker images, and show these images to a pathologist.

The next step is the extraction of information from the partitions such as the area, the average gray-scale or the greatest diagonal. Then the analysis of these data will be possible.
Chapter 5

Appendix

5.1 Could we use another process?

This report explained a way to segment the images. This section will explain that some other possibilities could be possible, and why we chose this one.

5.1.1 Our image processing

This is a quick review of our algorithm:

![Image processing algorithm](image)

Figure 5.1: Our algorithm. A more precise graph is shown in the beginning of the chapter.

5.1.2 Other ways to segment the image

Another way to segment the image, that seems \textit{a priori} better is to keep dividing the regions:

1. background segmentation;
2. cell segmentation (boundary of the cytoplasm);
3. nucleus segmentation.

The main problem is that the extraction of the cytoplasm boundaries are guessed and not truly determined. This segmentation is made impossible by the fact that you can not find the separation of two cells without searching between two nuclei.

Using the reverse segmentation (nucleus, cytoplasm, background) could be possible. What lead us to find the background at first is the following reasons:

- it is easy to find and we do not need any further informations;
- it seems that the nucleus extraction can be done without knowing the background; but in fact the position extraction is based on thresholding itself based on heap information.
- the background occupies a large amount of space. Being able to ignore it can speed up the segmentation of other objects.

In particular, if no cells are found on the image, everything can be skipped.

5.2 Algorithm of the Ph. D of Lezoray (2000)

In Lezoray et al. (2000a), the use of the color information enable him to have better contrasts.

Figure 29. Graphe de la segmentation des cytoplasmes.

Figure 5.2: Cytoplasm segmentation in Lezoray et al. (2000a)

1This is true for our segmentation, but it is true that we could get rid of this information
Figure 32. Graphe de la segmentation des noyaux.

Figure 5.3: Nuclei segmentation in Lezoray et al. (2000a)
Bibliography


