V6 - Digital pathology and MRI diagnostics

Pathology (from the Greek roots of *pathos* (πάθος), meaning "experience" or "suffering", and *-logia* (-λογία), "study of") is an important part of the causal study of diseases and a major field in modern medicine and diagnosis.

Digital pathology (DP) includes all aspects of

- acquisition,
- process management, and
- data interpretation

to yield pathology information from a digitized pathology sample's image.

Program for today

→ **Q1: Is the tissue healthy or cancer?**

→ **Q2: What cancer is it?**

→ **Q3: Where is the cancer?**

www.wikipedia.org
Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

V7 Processing of Biological Data WS 2021/22 1

In lecture 6, we will discuss some aspects arising when processing of imaging data with relation to bioinformatics.

Recently, during the past 7-8 years or so, more and more bioinformatics groups have started to engage here.

This is particularly due to the enormous success of applying Deep Learning methods to image analysis.

Of course, we can only touch on certain aspects here.

Those of you who want to go deeper into the field of imaging data may want to consider attending

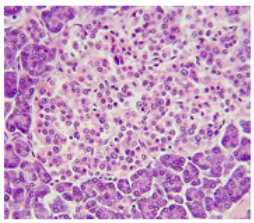
e.g. the core lecture IPCV of Prof. Joachim Weickert,

https://www.mia.uni-saarland.de/Teaching/ipcv21.shtml

Include biological or chemical markers or tissues

Staining tissues with **hematoxylin and eosin** (H&E) involves application of hemalum, a complex formed from aluminum ions and hematein.

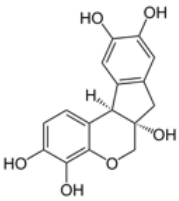
Hemalum colors nuclei of cells (and a few other objects) **blue**. The nuclear staining is followed by counterstaining with an aqueous or alcoholic solution of **eosin Y**. This solution colors eosinophilic structures in various shades of **red**, **pink** and **orange**.



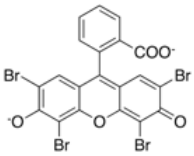
Alternatives to H&E staining are:

- Immunohistochemical (IHC) imaging
- label-free methods based on spectral imaging.
- Direct recording of chemical composition

  → no need for dyes or stains.



Hematoxylin



eosin Y

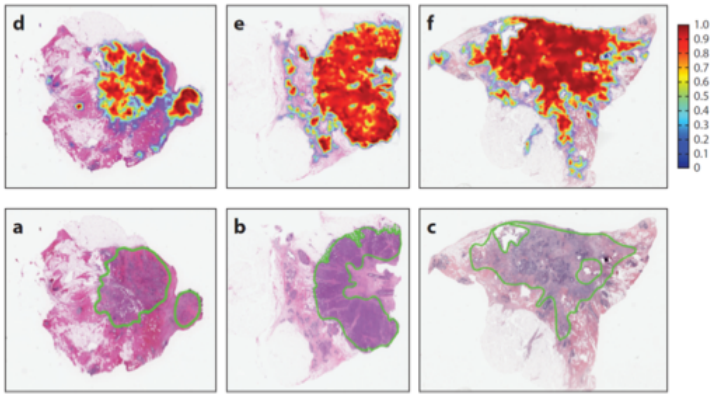
www.wikipedia.org
Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

Histology stands for studying the microanatomy of cells, tissues, and organs as seen through a microscope.

In traditional histology done by pathologists at hospitals or medical laboratories, the biopsies taken from patients are „stained“ (colored) by application of chemicals.

Digital pathology: where is the tumor?

(*Top*) Probabilistic output of a deep learning classifier for regions of invasion.



(*Bottom*) Corresponding hematoxylin and eosin images with a pathologist's markup of the extent of cancer extent.

Note the concordance between the two rows of images.

Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412
(its Fig. 1 has mixed-up legend legend – correct in *Proc. SPIE* 9041:904103)

The exact identification of tumor regions in biopsy images is a very difficult, tedious and important job for pathologists.

Can this also be done equally well by a computer algorithm?

The top row shows the regions identified by a layered neural network. The color intensities denote the level of confidence.

The bottom row shows the outcome when the same task is done by a pathologist. The tumor region detected by him/her is marked by a green line.

This is a link to the review paper cited:
https://www.annualreviews.org/doi/abs/10.1146/annurev-bioeng-112415-114722

This is a link to the original publication:
https://engineering.case.edu/centers/ccipd/sites/ccipd.case.edu/files/Automatic_detection_of_invasive_ductal_carcinoma_in_whole.pdf

Quantitative histomorphometry

Quantitative histomorphometry (QH) involves computerized image analysis tools for quantitatively assessing cancer tissue and non–cancer tissue morphology and architecture.

QH measurements can be divided broadly into 3 groups:

(a) architectural,

(b) shape, and

(c) texture based.

The online version of the dictionary by Merriam-Webster provides as a **Medical Definition of *histomorphometry*:**

the quantitative study of the microscopic organization and structure of a tissue (as bone) especially by computer-assisted analysis of images formed by a microscope.

Listed here are 3 types of measurements used in quantitative histomorphometry.

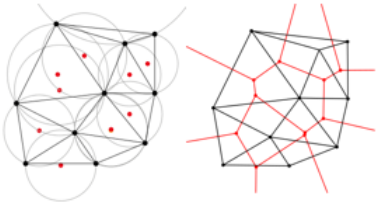
(a) Architectural QH measurements

Architectural features capture the arrangement and **spatial topology** of histologic primitives such as individual nuclei, tubules, mitoses, and lymphocytes.

The spatial location of a particular primitive is considered to be a node in a graph.

The nodes are then connected using graph construction algorithms [e.g., Voronoi diagram, Delaunay triangulation, minimum spanning tree].

The **Delaunay triangulation** with all the circumcircles and their centers (in red).



Connecting the centers of the circumcircles produces the **Voronoi diagram** (in red).

Quantitative measurements (e.g., inter-node distance, clustering coefficient of the nodes = density of links between the neighbors of a node) can quantitatively characterize the graph and, hence, the image.

www.wikipedia.org

V7 Processing of Biological Data WS 2021/22 5

The strategy of architectural measurements has traditionally been used to classify detectable objects (here called: primitives) in images of biological cells.

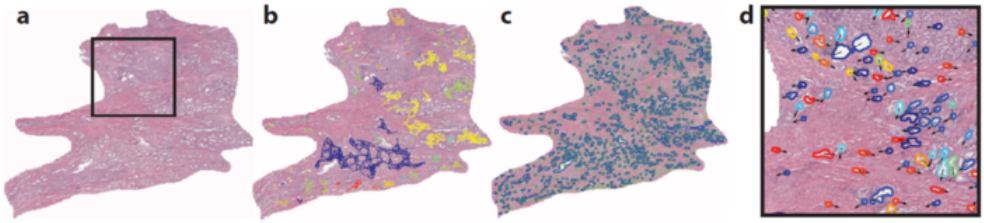
For example, one measures distances between objects.

Delaunay triangulation is named after Boris Delaunay, a Russian mathematician who published 1934 on „la sphere vide“ (the empty sphere), http://www.mathnet.ru/links/a5622f49bf1d96669a7aeacc0ab1d3f6/im4937.pdf

The triangles connect (some of) the data points so that to other data point lies in the circumcircle through the corners of a triangle.

This concept is similar to the concept of molecular fingerprints that is used in chemoinformatics in order to characterize molecules.

(a) Global and cell cluster graphs



(*a*) Prostate cancer tumor region. The region of interest (ROI) is outlined in blue.
(*b*) Cluster graphs establish localized **gland networks**.
(*c*) Delaunay triangulation reveals a global graph which traverses stromal and epithelial boundaries, whereas co-occurring gland tensors compute localized features from the gland networks.
(*d*) The ROI from panel *a*. The color map of the gland orientations (**red** is 0° , **blue** is 180°) demonstrates the variation in local gland orientation.

Gland orientations are architecturally differently arranged in tissue from patients with and without disease recurrence.

Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

In German, a „gland“ means „Drüse“. „Gland networks“ would then be compact clusters of glands.

In panel (d), the red and blue colors mark glands oriented toward opposite directions.

Interestingly, patients with disease recurrence (most likely tumors …) were observed to have different gland orientations than patients who were cured.

This suggests that characterizing the gland topology could allow making predictions about the likelihood that the disease may come back (recur).

(b) Shape QH measurements

The **shape** of individual histologic primitives can indicate the presence of disease.

Shape features such as

- fractal dimension: ratio comparing how a detail in a pattern changes with the scale at which it is measured
- angularity, size, and
- smoothness of the boundary

were found to differ between nuclei and glands in high and low grades of prostate and breast cancers.

As the length of the measuring stick is scaled smaller and smaller, the total length of the coastline measured increases (-> fractal dimension)

Also, the disorder (or entropy) in the orientation of nuclei and glands in prostate tissue was related to the tumor recurrence in patients with prostate cancer.

www.wikipedia.org

On the previous slide, we heard that gland orientations may be related to disease recurrence.

Also, certain **shape features** were reported to show characteristic differences for differently severe grades of prostate and breast cancer.

(c) Texture-based QH measurements

Texture refers to quantitative measures of **spatial neighborhood interactions** between pixel intensities within local neighborhoods in an image.

These could include

- first-order spatial intensity interactions (e.g., mean, standard deviation, median, variance) within local neighborhoods and
- second-order interactions (e.g., co-occurrence features).

More complex textural features can also be extracted; these include steerable and multiscale gradient features via mathematical operators such as Gabor filters, local binary patterns, and Laws filters.

The shape and texture of nuclei within the stroma are significantly correlated with disease recurrence and patient outcome in breast, prostate, and oropharyngeal cancers.

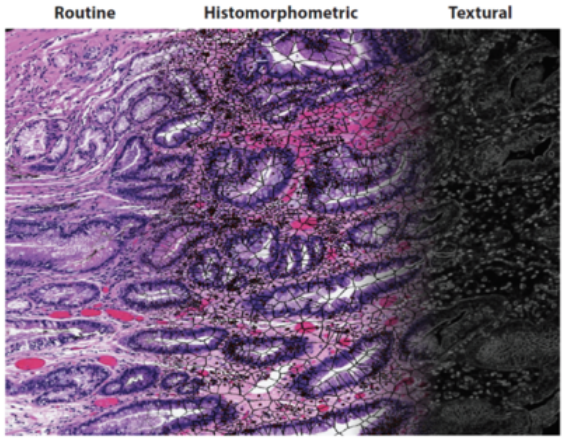
„**Texture**“ stands for the feel, appearance, or consistency of a surface or a substance, e.g. "skin texture and tone"

Here, one characterizes textural features of tissue.

(c) a digital stain

(Left) A routine hematoxylin and eosin tissue image.

The left image can be converted into a histomorphometric representation comprising nuclear architecture (*middle*) and textural measurements (*right*).



Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

The figure shows the digital stain representation of a routine H&E image (left), with overlays of nuclear architecture networks (middle) and capture of stromal and epithelial textural variations (right).

Principles of chemical imaging

IR imaging provides high image contrast, fast data recording, and high molecular sensitivity.

Vibrational frequencies within molecules directly resonate with optical frequencies in the mid-IR spectral region.

Thus, light absorption provides a quantitative **molecular fingerprint** of the material, providing ample molecular biomarkers.

No dyes or stains are needed to visualize molecular content.

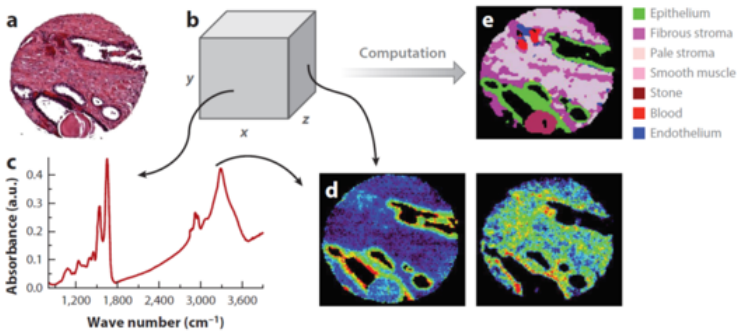
Data can be recorded without prior knowledge of the type or composition of the sample.

There are techniques that try **to combine different sorts of data acquisition**, e.g. imaging the sample by visible light AND by infrared (IR) spectroscopy.

The idea of this is to annotate the molecular composition of the objects detected by light imaging.

Chemical imaging

(*a*) Conventional imaging in pathology requires dyes and a human to recognize cells.



(*b*) In chemical imaging data, both (*c*) a spectrum at any pixel and the spatial distribution of any spectral feature can be observed, as in (*d*, *left*) nucleic acids (at ~1,080 cm^{-1}) and (*right*) collagen (at ~1,245 cm^{-1}).

(*e*) Computational tools can then translate the chemical imaging data into knowledge used in pathology.

Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

V7 Processing of Biological Data WS 2021/22 11

In (d), one records the IR spectra with spatial resolution.

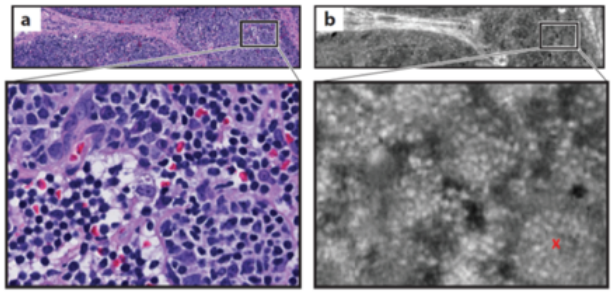
Certain wavenumbers (e.g. 1080 cm-1 and 1245 cm-1) are characteristic for certain molecules that can be assigned to cellular compartments.

E.g. nucleic acids are located in the **nucleus**. Collagen is the main structural protein in the **extracellular matrix** in the various connective tissues in the body.

If the composition of various tissues is known (see panel (e)), one can assign the tissue type to the detected continuous regions by a suitable software.

Comparison of H&E stain and IR imaging

Comparison of hematoxylin and eosin (H&E)-stained optical microscopy and infrared (IR) images of lymph node tissue.



There is a slight discordance between the H&E and IR images because they are recorded for different tissue sections.

(*a*) An H&E-stained image from a healthy lymph node biopsy.

(*b*) A high-definition IR image of a serial section of the lymphoid tissue.

Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

(Left) Optical (light) microscopy image of lymph node tissue.

(Right) IR image of a related section of the same tissue.

The upper row are images at lower resolution.

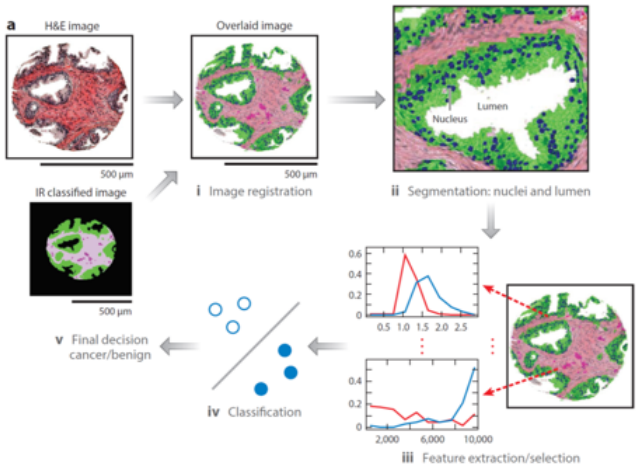
The bottom row shows a blowup of the rectangular region marked in the upper row.

Both types of spectroscopy/microscopy reveal different details of the same sample.

Overview of multimodal digital pathology system

(*i*) A Fourier transform infrared spectroscopy data–based cell type classification is overlaid on a hematoxylin & eosin -stained image, -> (ii)

(*ii*) **segmentation** of nuclei and lumen in the tissue sample.



then (*iv*) used by the classifier to (*v*) predict whether the sample is cancerous or benign.

(*iii*) Features are extracted and selected,

Bhargava, Madabhushi
Annu. Rev. Biomed. Eng. 2016. 18:387–412

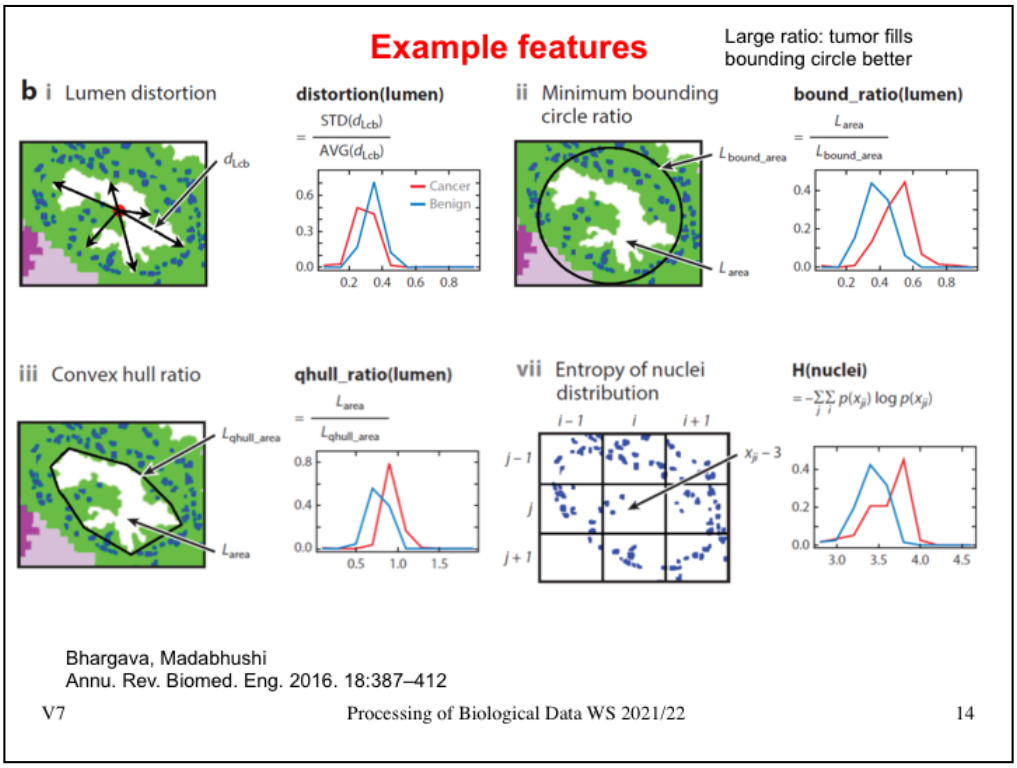
V7 Processing of Biological Data WS 2021/22 13

You should read this slide in clockwise direction starting at the top left.

In biology, a lumen (plural lumina) is the inside space of a tubular structure, such as an artery or intestine.

In cell biology, a lumen is a membrane-defined space that is found inside several organelles, cellular components, or structures: thylakoid, endoplasmic reticulum, Golgi apparatus, lysosome, mitochondrion, or microtubule.

Here, we are dealing with larger lumenal volumes outside of cells. Note the scale bar (500 µm). Single cells have dimensions of a few µm. Hence, the many blue circles in the right plot are separate nuclei of separate cells.



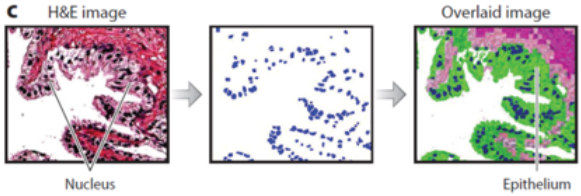
This example was taken from the following study:
https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-11-62

There, the authors mention that the characteristics of nuclei and lumens change in cancerous tissues. In H&E stained images, lumens are recognized to be empty white spaces surrounded by epithelial cells. In normal tissues, lumens are larger in diameter and can have a variety of shapes. In cancerous tissues, lumens are progressively smaller with increasing grade and generally have less distorted elliptical or circular shapes.

Classification of tumor tissue

IR and H&E images can be overlaid with an automated alignment algorithm.

The features allow better classification of cancer than does H&E staining alone.



AUC, area under the curve;

AVG, average;

STD, standard deviation

	10 CV (Data 1)		10 CV (Data 2)		Validation	
	AUC		AUC		AUC	
	AVG	STD	AVG	STD	AVG	STD
IR and H&E	0.982	0.0030	0.974	0.0145	0.956	0.0089
H&E only	0.968	0.0052	0.880	0.0175	0.918	0.0100

10 CV: 10-fold cross validation

Classification of prostate samples works very well.

"Validation": SVM, trained on Data1 and applied to Data2.

V7 Processing of Biological Data WS 2021/22 15

https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-11-62

Data set 1: 66 benign tissue samples and 115 cancer tissue samples.

Data set 2: 14 benign and 36 cancer tissue samples.

The aim was to distinguish cancerous from non-cancerous tissue samples using a trained support vector machine.

Case study: classification of lung cancer from raw images

nature medicine ARTICLES

Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning

2 most prevalent types of lung cancer:

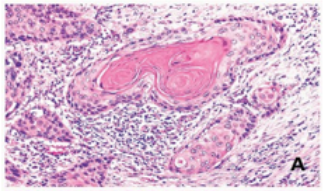
LUSC – lung squamous cell carcinoma (SCC):

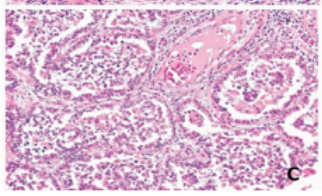
SCCs are different types of cancer that result from **squamous cells** (type of epithelial cell).

LUAD – lung adenocarcinoma - adenocarcinoma forms in mucus-secreting **glands** throughout the body. It can occur in many different places in the body.

Both are **non-small cell lung cancers**

SCC





AD - BAC

Coudray et al. Nature Medicine 24, 1559–1567 (2018)
http://www2.keelpno.gr/blog/?p=1391

Link to the paper: https://www.nature.com/articles/s41591-018-0177-5

Deep Learning methods are nowadays known to be highly successful in automated classification of biomedical images.

This is only one example from an amazing mass of similar publications.

Here, the authors tried to distinguish two types of non-small cell lung cancers, LUSC (top figure) and LUAD (bottom figure).

Since Sept. 2018, this paper has been cited about 1000 times.

Treatment of LUAC / LUSC

Stage I – surgery or radiation therapy

Stage II – surgery and chemotherapy or radiation therapy

Stage III – sequential or concurrent chemotherapy and radiation therapy, more options …

Stage IV – patient genetics becomes important

- Cytotoxic combination chemotherapy
- Combination chemotherapy with monoclonal antibodies
- Maintenance therapy after first-line chemotherapy (for patients with stable or responding disease after 4 cycles of platinum-based combination chemotherapy)
- EGFR tyrosine kinase inhibitors
- ALK inhibitors (for patients with ALK translocations)
- ROS1 inhibitors (for patients with ROS1 rearrangements)
- BRAFV600E and MEK inhibitors (for patients with BRAFV600E mutations)
- Immune checkpoint inhibitors with or without chemotherapy

https://www.cancer.gov/types/lung/hp/non-small-cell-lung-treatment-pdq#section/_48406

These are the treatment options at various stages of LUAC / LUSC, see table 7 at the given website of the National Cancer Institute of the US.

Further down, the stages become more serious, and the treatment more aggressive.

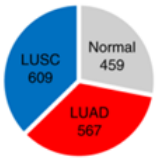
It is very important to identify in detail what type of tumor a patient has.

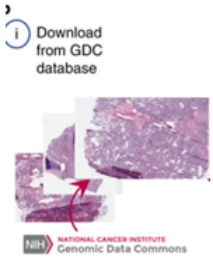
To select the best treatment in stage IV, it is also important to find out if the patient (or even the tumor tissue) carries particularly genomic mutations.

Classification of tumor tissue

Q: Can one classify LUAD / LUSC / normal (healty) by deep learning at similar accuracy as a medical expert (pathologist)?

Use tumor slides from TCGA (The Cancer Genome Atlas):





Coudray et al. Nature Medicine 24, 1559–1567 (2018)

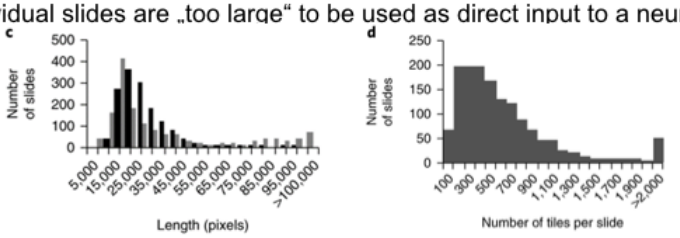
The authors used tissue images (bottom figure) provided on the TCGA website.

As shown in the top figure, TCGA provides roughly equal portions of LUSC, LUAD, and normal (healthy).

But these numbers seem not enough for application of Deep Learning methods which require massive amounts of training data (100000s) in order to train a successful classifier.

Classification of tumor tissue

Individual slides are „too large“ to be used as direct input to a neural network.



Idea: split each slide into „tiles“ of 512 × 512 pixels.

This largely increases the amount of training data.

Split data into 70% for training, 15% for validation, and 15% for testing.

Remove tiles where > 50% of the surface is covered by background (too dim).

-> **about 1 million tiles**

Coudray et al. Nature Medicine 24, 1559–1567 (2018)

Here, there are only 400 – 600 images each. The solution found by the authors was very simple.

They argued that the available images are „too large“ = have too high resolution to be used directly.

So one option would be to use only small parts of each image. But then, the amount of data is too small.

Also one would throw away a lot of potentially useful data.

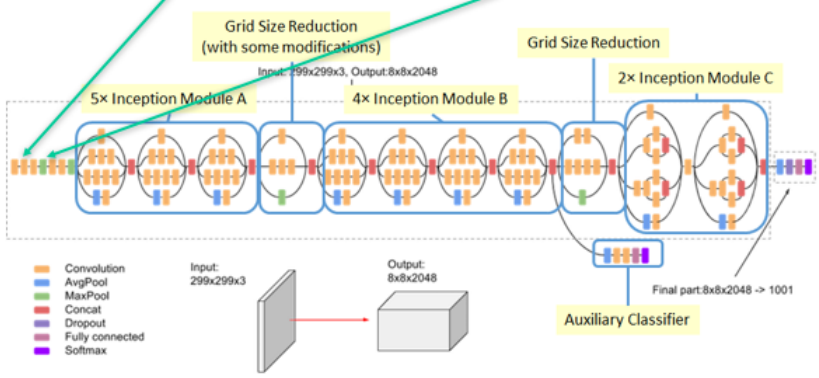
Therefore, the idea was to split each slide into many small images and assume that the contained information is not largely redundant.

This increased the amount of data to more than 1 million smaller images (tiles).

Deep learning model

The authors used a convolutional neural network architecture invented by Google that is termed inception v3 architecture36:

5 initial convolution nodes are combined with 2 max pooling operations and followed by 11 stacks of inception modules



Implementation with `TensorFlow` software by Google.

medium.com

This is a brief introduction of the type of Deep Learning neural network architecture used in this study.

„Convolution“ nodes integrate density information from a region into a central pixel. See

https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c

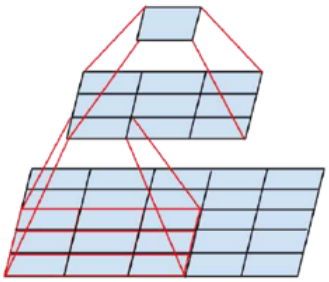
for more information about this particular architecture (inception v3 architecture36).

inception v3 architecture36

The idea of staggered convolution nodes (also called: factorization into smaller convolutions) is to reduce the number of parameters that need to be trained.

In the example below, two 3×3 convolutions replaces one 5×5 convolution

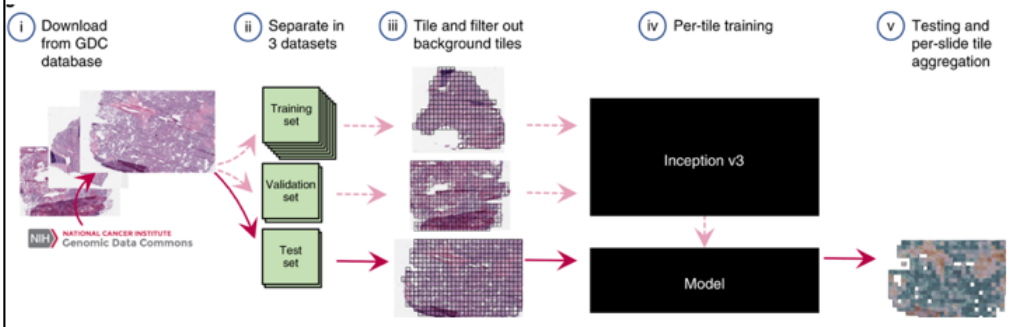
By using 1 layer of 5×5 filter, number of parameters = $5 \times 5=25$
By using 2 layers of 3×3 filters, number of parameters = $3 \times 3+3 \times 3=18$
-> The number of parameters is reduced by 28%



medium.com

Principles of this particular architecture (inception v3 architecture36).

Workflow



Classification of normal versus tumor tissues (~0.99 AUC) and distinguishing lung cancer types can be done with high accuracy (0.97 AUC).

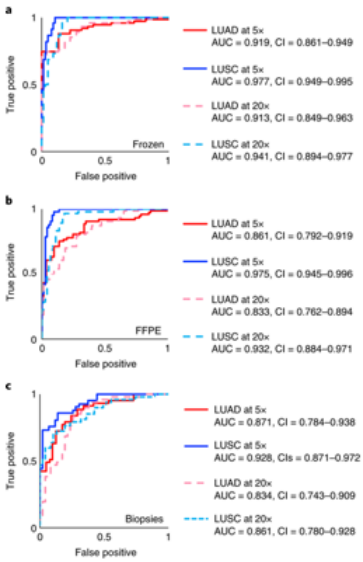
This is the same accuracy as observed for 3 trained pathologists who were asked to classify the same data.

Coudray et al. Nature Medicine 24, 1559–1567 (2018)

It has been reported in recent years that deep learning methods can be extremely successful in image recognition.

Many companies (also pharma companies) have now opened deep learning groups who work on such tasks.

Classify presence and type of tumor in alternative cohorts



Use the trained model on alternative cohorts.

→ Check robustness.

(**a**) Receiver operating characteristic (ROC) curves from tests on **frozen sections** (*n* = 98 biologically independent slides)

(**b**) **FFPE** (Formalin-Fixed Paraffin-Embedded) **sections** (*n* = 140 biologically independent slides)

(**c**) **biopsies** (*n* = 102 biologically independent slides.

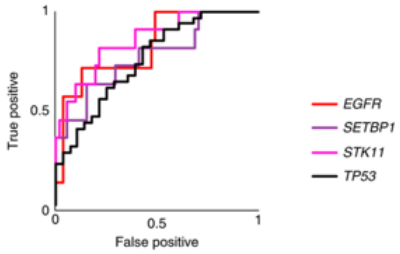
5x optical magnifications give better results than 20x.

Coudray et al. Nature Medicine 24, 1559–1567 (2018)

Here, the authors demonstrate that the trained classifier does not only perform well for the dataset on which it was trained, but also for 3 alternative cohorts.

Classification of genetic variants

Can CNNs be trained to predict gene mutations using images as the only input?



Somehow. The accuracy (AUC) is between 0.64 (LRP1B) and 0.84 (STK11).

Even better results can be expected when more training data becomes available.

Tool may be helpful to assist pathologists in their routine work.

Coudray et al. Nature Medicine 24, 1559–1567 (2018)

V7 Processing of Biological Data WS 2021/22 24

Before this study, it was unclear whether gene mutations would affect the pattern of tumor cells on a lung cancer whole-slide image.

Interestingly, training the network using the presence or absence of mutated genes as a label revealed that there are certain genes whose mutational status can be predicted from image data alone: *EGFR*, *STK11*, *FAT1*, *SETBP1*, *KRAS*, and *TP53*. The ability to quickly and inexpensively predict both the type of cancer and the gene mutations from histopathology images could be beneficial to the treatment of patients with cancer given the importance and impact of these mutations

Q2: where is the tumor? Example: Wilms tumor

Wilms tumor, also known as **nephroblastoma**, is a cancer of the kidneys that typically occurs in children, rarely in adults.

It is named after Dr. Max Wilms, a German surgeon (1867–1918) who first described it.

Approximately 500 cases are diagnosed in the U.S. annually (rare tumor).

The majority (75%) occur in otherwise normal children; a minority (25%) are associated with other developmental abnormalities.

Wilms tumor is highly responsive to treatment, with about 90% of patients surviving at least five years.

Diagnose tumor e.g. with MRI scan:

This is a sort of NMR experiment.

Measure T1 and T2 spin relaxation times of tissues.

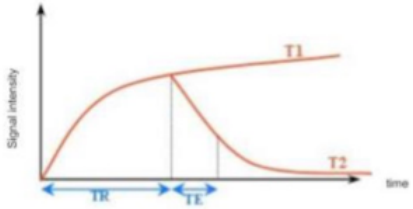


Figure 1: T1 and T2 weighting times [R21]

In the second half of this lecture, we discuss again the problem of detecting a tumor region.

This part was taken from the MSc thesis in computer science of Vera Bazhenova (supervised by me).

The project was motivated by the work of Prof. Norbert Graf / UdS medical school on a child tumor termed Wilms tumor.

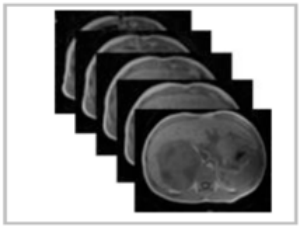
Fortunately, this is a rather rare tumor.

MRI stands for magnetic resonance imaging. You can read more about the MRI technique at https://casemed.case.edu/clerkships/neurology/Web%20Neurorad/MRI%20Basics.htm

I cite from that site „Tissue can be characterized by two different relaxation times - T1 and T2. T1 (longitudinal relaxation time) is the time constant which determines the rate at which excited protons return to equilibrium. It is a measure of the time taken for spinning protons to realign with the external magnetic field. T2 (transverse relaxation time) is the time constant which determines the rate at which excited protons reach equilibrium or go out of phase with each other. It is a measure of the time taken for spinning protons to lose phase coherence among the nuclei spinning perpendicular to the main field."

Non-invasive MRI diagnostics: data sets

Vera Bazhenova (MSc Comp Sci UdS 2014) analyzed vertical cross section MRI sets of scans for patients with nephroblastoma tumor.



Each set contains 20 to 50 scans.

The following part of this lecture was taken from her MSc thesis.

Aim of this project:

Identify precise **location** of the tumor.

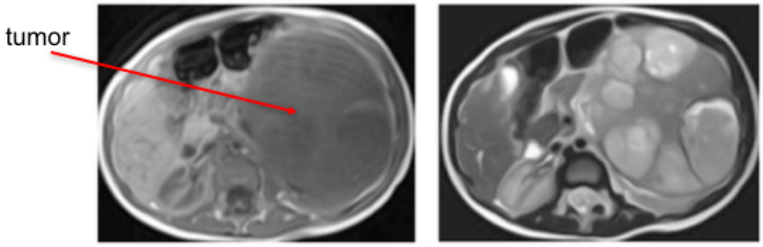
This can be basis for surgery (where to operate?) or be used for diagnostic purposes (follow tumor growth).

V7 Processing of Biological Data WS 2021/22 26

The available data provided by Prof. Graf consisted of vertical cross sections from MRI scans that characterize the body signals at different body height.

Input data

T1-weighted scan T2-weighted scan



T1-weighted scans appeared more suitable for digital analysis since the tumor region has a more homogeneous contrast.

The body contours are well visible and can be easily distinguished from the background.

V7 Processing of Biological Data WS 2021/22 27

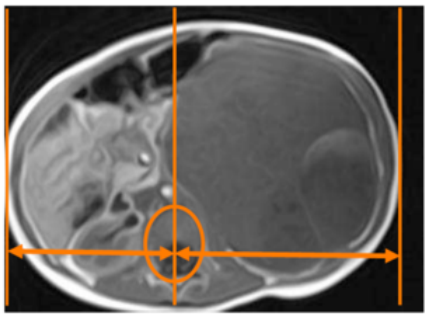
This is a vertical slice through a child‘s body. An arrow marks a very large region that takes up almost half of the slice area. This is – very sadly – a gigantic nephroblastoma tumor of the right kidney.

Use spine location to detect asymmetry

The nephroblastoma tumor affects in more than 95% of the cases only one kidney of the patient.

In healthy individuals, the spline is located in the center of the body cross section.

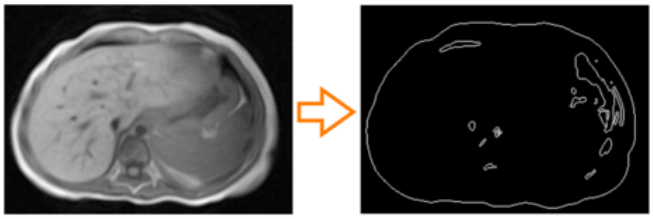
When the affected kidney grows abnormally, the spine appears shifted either to the left or to the right side.



In order to locate the tumor position, we found it helpful to use the spine as reference point.

For this patient, the spine (marked by arrows and circled) appears shifted toward the left side, simply because the tumor has grown so large.

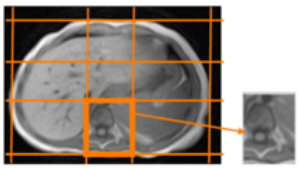
Determine perimeter



To locate the spine region, the body boundary is detected using a **perimeter detection function** applied to a binary image.

A pixel is considered as a part of the perimeter if it has a nonzero brightness and it is connected to at least one zero-valued pixel.

The “region of interest” for the spine is vertically located in the middle third and horizontally in the lower third of the body.



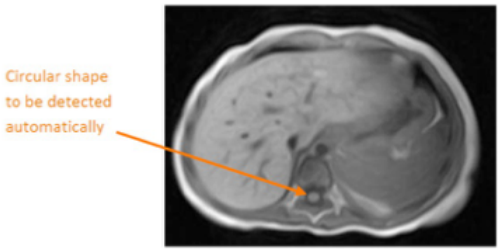
V7 Processing of Biological Data WS 2021/22 29

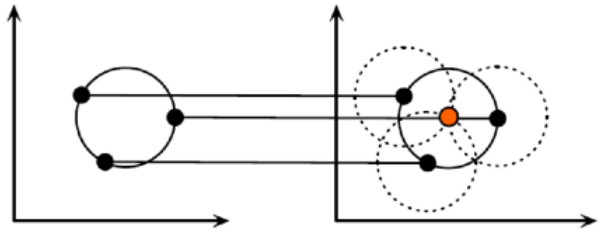
In this application scenario, pixels with zero-values occur only outside the body.

Task: automatic detection of spine

In a T1-weighted MRI scan, the middle of the spine M_s appears as a white circle at the level of the liver.

→ Apply the **circular Hough transform** to the first scans of a series until a spine center is detected.





Each point in geometric space (left) generates a circle in parameter space (right). The circles in parameter space intersect at the (a, b) that is the center in geometric space.

https://www.cis.rit.edu/class/simg782/lectures/lecture_10/lec782_05_10.pdf

V7 Processing of Biological Data WS 2021/22 30

A circle with radius R and center (a,b) can be described with the parametric equations

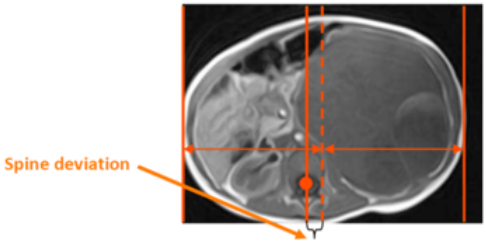
$x = a + R*\cos(\theta)$

$y = b + R*\sin(\theta)$

In this case, the objective is to find the (a,b) coordinates of the spine center.

The Hough transform can be used to determine the parameters of a circle (coordinates of center and radius) when a number of points that fall on the perimeter are known.

Spine position



Detect the spine middle in all scans of the MRI series.

A patient who shows a significant **deviation** of the spine from the center is flagged as candidate to have a certain class of diseases including a nephroblastoma tumor.

The **direction** of the deviation indicates to us which side of the body is likely affected by this disease.

The position of the spine center relative to the x-axis midpoint of the perimeter is used to classify which kidney may be affected.

Masked scan

If a disease is present, we prepare a **body mask** that hides

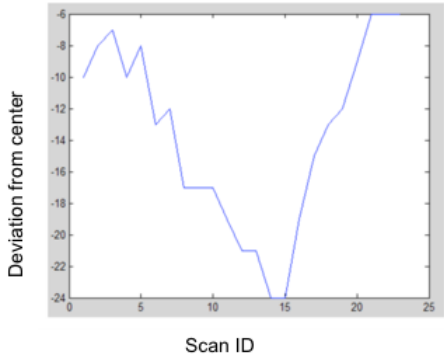
- the spine (1),
- the region below the spine (2),
- the body perimeter (2) and
- the side which presumably does not contain a tumor (3).

Masking of the non-affected regions helps to simplify subsequent tasks.

Spine deviation curve

Another output of the spine detection algorithm is the index of the scan with the **maximum deviation** of the spine from the center.

This index is used in order to extract the gray value range of the tumor in order to enhance the accuracy of the tumor recognition algorithm.



The figure shows the spine deviation curve for a real MRI scan.

According to the coordinate system adopted here, a negative deviation means that a disease occurs on the right side of the body

In order to localize the tumor, the algorithm needs to know its brightness (gray value).

This gray value level can be best characterized in the scan showing the largest portion of the tumor.

The assymetric spine position may be helpful to identify this scan.

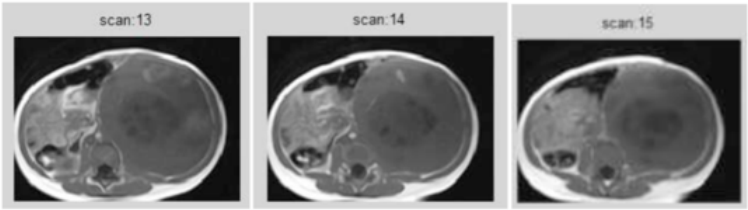
In the example shown here, scan #14 shows the largest deviation from the center (in millimeters?)

Tumor detection

Detection of the tumor is performed in two main steps.

In the first step, the tumor gray value range is determined.

In the second step, the precise region of the tumor is detected.

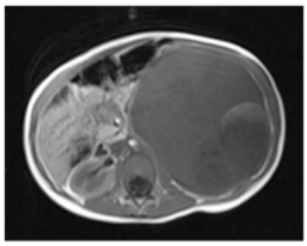


- Use the scan with the largest deviation
- Identify the largest blob. Even if the liver is on the same side as the tumor, the tumor is likely already larger than the liver.

From experience, one knows that the largest blob (Merriam Webster: a spot of color) is typically the tumor.

Image denoising

The delivered MRI scan series are usually quite noisy and need to be pre-processed in order to be suitable for detecting the tumor.



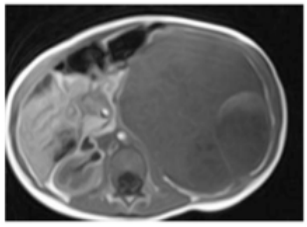
a. Original MRI scan

For this, **diffusion filtering** is used. This denoising algorithm removes noise while it preserves edges.

$$\frac{\partial \rho(\vec{r},t)}{\partial t} = -\nabla(-D\,\nabla\rho(\vec{r},t)) = D\,\Delta\rho(\vec{r},t)$$

Diffusion equation.

The above "diffusion equation" is applied iteratively to an input image until the output becomes smooth enough and reaches the wished noise elimination.



b. Filtered MRI scan

In addition, other filters are applied, e.g. the median filter

V7 Processing of Biological Data WS 2021/22 35

The diffusion equation describes how a region of larger density leaks out into regions of lower density over time.

Here, the concentration is described by variable ρ (rho).

The diffusion equation describes how the concentration at location $\boldsymbol{r}$ and time t changes over time. Its first time derivative is equal to the second spatial derivative times the diffusion coefficient.

This means that if there is higher density everywhere around a particular grid point, its second spatial derivative of the density is positive (curved upwards).

Thus, the concentration at this grid point will increase over time as long as the concentration at this point reaches that of neighboring points.

The diffusion equation is a very helpful tool to remove „noise" (brightness fluctuations) from an image.

Small excursion: continuity equation

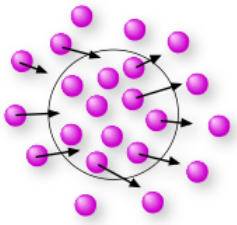
We can derive the diffusion equation quite simply:

1) Continuity equation asks where does the material go?

$$\frac{\partial \rho(\vec{r},t)}{\partial t} = -\nabla \vec{j}(\vec{r},t) = -\text{div}\ \vec{j}(\vec{r},t)$$

Change of density ρ in (r, t) | Divergence of the current | = | Sources and sinks of the particles

partial derivative:
=> Consider only changes of ρ in a small time interval at the given position



$\Delta N = N_{in} - N_{out} = 3 - 5 = -2$

V7 Processing of Biological Data WS 2021/22 3

For a better understanding of the diffusion equation, we will quickly derive it by combining the continuity equation and Fick's first law.

The **continuity equation** describes the transport of some quantity.

In the bottom example, three purple particles would enter into the circled area during a given time interval from the left side.

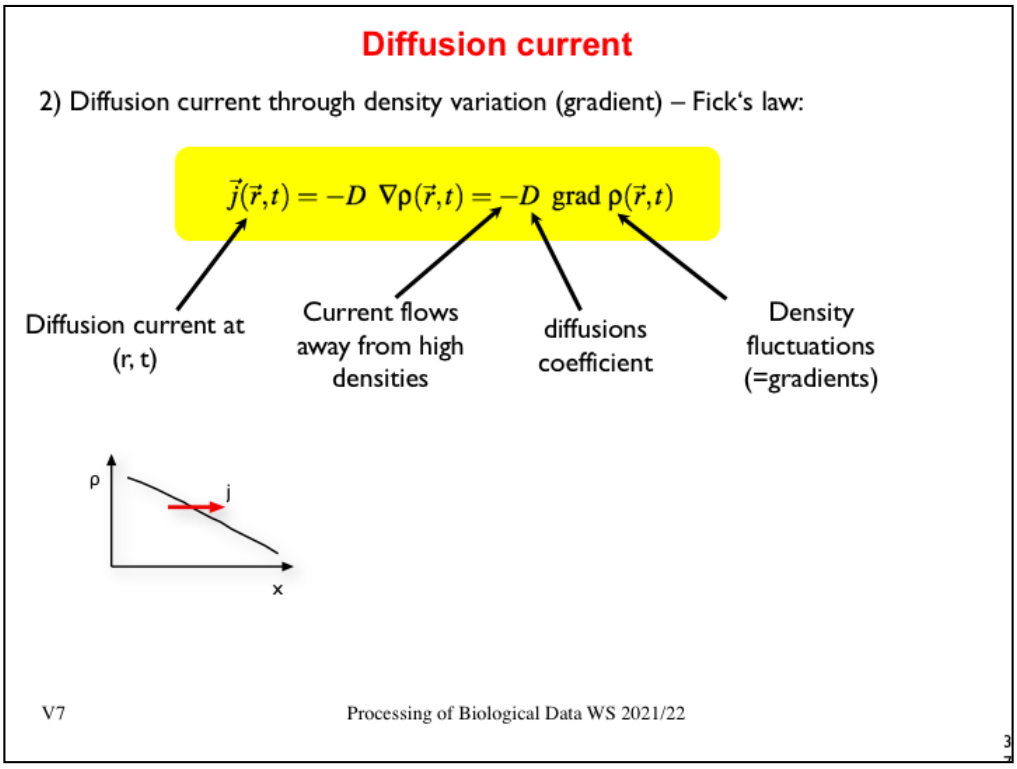
In the same interval, five particles would leave the circle to the right. Thus, the net number of particles in the circle reduces by 2.

The current (flux) j increases from left to right. Hence, its first spatial derivative (divergence) is positive. According to the continuity equation, the concentration in the circle decreases over time.

This is an example how we can obtain the change of density by comparing the magnitude of the particle current between left and right sides, which is described by the continuity equation.

Latex code for the continuity equation:

```
\frac{\partial \rho(\vec{r}, t)}{\partial t} = - \nabla
\vec{j}(\vec{r}, t) = - \mbox{div }\vec{j}(\vec{r}, t)
```



The second equation we need is Fick's law.

It describes in which direction a diffusion current will be directed. Matter always flows from a high concentration region to lower concentration regions. The gradient gives the direction in which the density increases most. Thus, the diffusion current will be directed into the opposite direction.

Latex code of diffusion current:

```
\vec{j}(\vec{r}, t) = -D \;\;\nabla \rho(\vec{r}, t) = -D
\;\;\mbox{grad } \rho(\vec{r}, t)
```

Derivation of diffusion equation (PDE)

Enter diffusion current

$$\vec{j}(\vec{r},t) = -D\ \nabla\rho(\vec{r},t) = -D\ \text{grad}\ \rho(\vec{r},t)$$

in continuity equation

$$\frac{\partial\rho(\vec{r},t)}{\partial t} = -\nabla\vec{j}(\vec{r},t) = -\text{div}\ \vec{j}(\vec{r},t)$$

=> diffusion equation:

$$\frac{\partial\rho(\vec{r},t)}{\partial t} = -\nabla(-D\,\nabla\rho(\vec{r},t)) \overset{D(\vec{r},t)=\text{const}}{=} D\,\Delta\rho(\vec{r},t)$$

$\Rightarrow$ The diffusion equation gives a complete description of the time- and space-dependent density (assuming no external forces such as gravity)

These 2 equations can now be simply combined to obtain the diffusion equation.

The diffusion constant can either be constant or variable. E.g. in a cell, the diffusion constant of proteins would be highest in the cytosol, but low in or near membranes.

The diffusion equation has become an important technique for image denoising.

Latex code:

```
\frac{\partial \rho(\vec{r}, t)}{\partial t} \;= \;\; - \nabla
(-D \;\nabla \rho(\vec{r}, t))\;\; =\; D\;\Delta \rho(\vec{r},
t)
```

```
D(\vec{r}, t) = \mbox{const}
```

FTCS–integrator

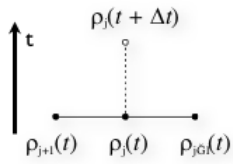
Diffusion equation with constant diffusion D in 1D:

$$\frac{\partial \rho(\vec{x},t)}{\partial t} = D\,\frac{\partial^2 \rho(\vec{x},t)}{\partial x^2}$$

Direct implementation on a lattice $\{\rho(x_i)\}$ with lattice spacing Δx :

$$\frac{\rho_j(t+\Delta t)-\rho_j(t)}{\Delta t} = D\,\frac{\rho_{j+1}(t)-2\rho_j(t)+\rho_{j-1}(t)}{\Delta x^2}$$

Forward in **T**ime — **C**entered in **S**pace



Propagation step:

$$\rho_j(t+\Delta t) = \rho_j(t) + \Delta t\, D\,\frac{\rho_{j+1}(t)-2\rho_j(t)+\rho_{j-1}(t)}{\Delta x^2}$$

Integration is stable for: $\Delta t \le \frac{\Delta x^2}{2D}$

Here, we introduce a simple algorithm termed „forward in time centered in space“ (FTCS).

For simplicity, we consider diffusion on a one dimensional lattice, where grid points have spacing Δx and are labeled by j.

In the middle equation, the left side is the discrete first derivative of the density at grid point j with respect to time.

The right side is the second spatial derivative of the density with respect to coordinate vector at position j.

More precisely, we consider how the first spatial derivative changes between right ($\rho_{j+1} - \rho_j$) and left ($\rho_j - \rho_{j-1}$) sides.

We then multiply the middle equation by Δt and add ρ_j (t) to both sides. This yields the integration algorithm for the propagation step.

This algorithm can easily be generalized to 3D.

The numerical algorithm is „stable“ (does not deviate from the analytical derivates too much) for the condition given at the bottom.

```
\frac{\partial \rho(\vec{x}, t)}{\partial t} =
D\;\;\frac{\partial^2 \rho(\vec{x}, t)}{\partial x^2}
\frac{\rho_j(t+\Delta t) - \rho_j(t)}{\Delta t} \;=\;
D\;\;\frac{\rho_{j+1}(t) - 2\rho_{j}(t) + \rho_{j-1}(t)}{\Delta
```

```
x^2}
\rho_j(t+\Delta t)  \;=\;  \rho_j(t)} +\Delta t\,
D\;\;\frac{\rho_{j+1}(t) - 2\rho_{j}(t) + \rho_{j-1}(t)}{\Delta x^2}
\Delta t \;\leq\; \frac{\Delta x^2}{2D}
```

Denoising: median filter

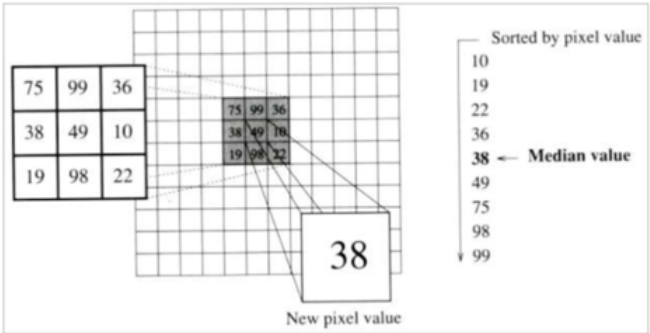


Figure 2: Median filter [R24]

With the median filter, the brightness of the central point is set to the median (38) of its direct neighbors.

Determine gray levels

Apply edge enhancement filter.

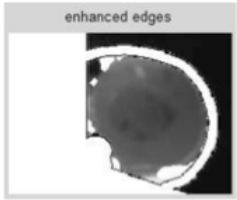
Then analyze the **histogram** of the resulting image.

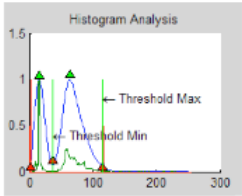
Extract minima and maxima in order to separate data clusters by applying the optimal thresholding.

Data clusters are then defined as maxima surrounded by minima.

The first cluster always represents the noise and the image background. The second cluster usually represents the tumor.

Hence the indices of the minima and maxima **(green bars)** of the tumor cluster should represent the gray value range of the tumor





grey value

The gray level of the tumor is determined from the brightness histogram.

The first cluster at lowest intensity contains noisy signals and the image background.

Experience suggests that the next peak belongs to the tumor tissue.

Fine detection of tumor blob

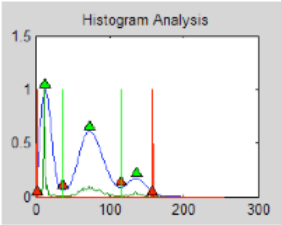
(1) Apply double thresholding using the just calculated threshold min and max gray values in order to extract the tumor blob.

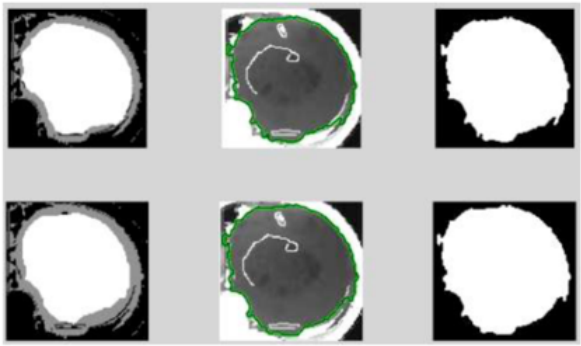
(2) fill the resulting image in order to get a mask.

(3) Subtracting this mask from the thresholded image gives us the body segmentation.

(4) Apply GrowCut on the extracted blob.

(5) Recompute histogram for this region.



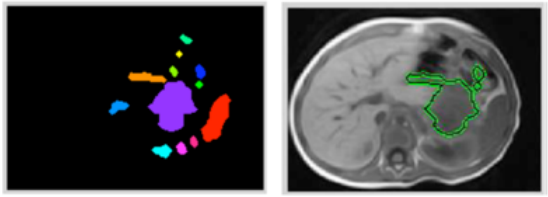


Using the min and max thresholds for the tumor gray values, the corresponding region in the image is cut out by applying a mask (middle column).

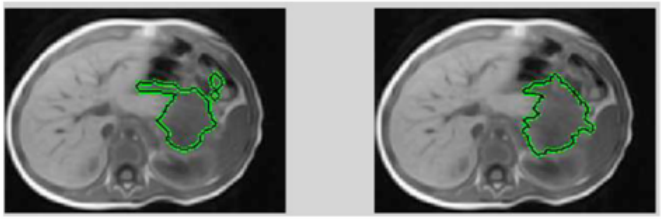
This yields the presumable tumor region shown in the right column.

This is a link to the paper describing the GrowCut algorithm: https://www.graphicon.ru/oldgr/en/publications/text/gc2005vk.pdf

Blob recognition: tumor detection



Apply some further hokus-pokus, e.g. blob detection



End result of automated tumor detection.

V7 Processing of Biological Data WS 2021/22 43

Vera Bazhenova applied further filters that she picked up from Prof. Weickert‘s lecture.

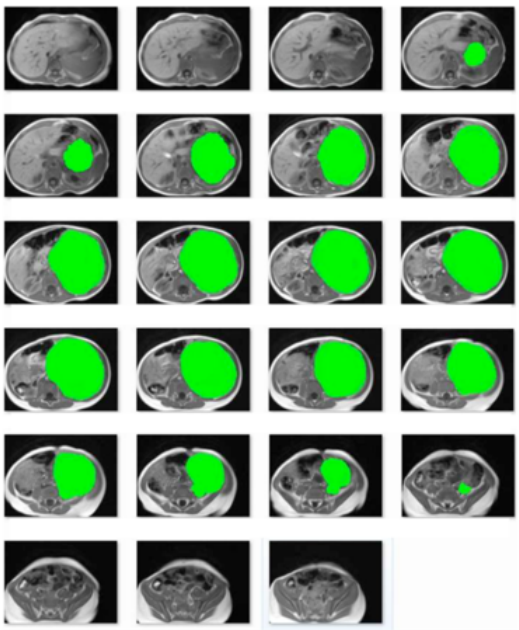
We will not go into the details of this here.

Gold standard

Gold standard:

Manually marked scans of series ID 2 from 1 till 20.

These are horizontal slices through the body at different levels from top to bottom.



V7 Processing of Biological Data WS 2021/22 44

To know how good the algorithm works, we need a gold-standard.

In this case, Vera Bazhenova tried her best and annotated the large region of homogeneous gray level manually herself.

Dimensions of tumor

„True“: defined by manual annotation

Scan ID	Mean Gray Level	Detected Area (mm²)	True Area	Perimeter	Centroid X	Centroid Y
4	72.95	1499.00	993.00	215.38	124.86	75.48
5	65.25	3142.00	1871.00	316.74	124.95	77.43
6	63.38	5120.00	4281.00	381.50	128.46	74.18
7	63.64	5876.00	5772.00	313.32	125.40	71.73
8	77.66	6511.00	6104.00	358.53	125.54	70.90
9	78.36	7509.00	6697.00	385.40	126.02	70.11
10	69.85	7887.00	6974.00	412.48	124.95	70.36
11	68.17	7782.00	6758.00	348.19	123.54	68.86
12	70.27	7922.00	6929.00	378.43	122.87	69.37
13	68.05	7779.00	7036.00	362.19	122.65	69.95
14	67.12	7253.00	6565.00	363.71	121.55	70.73
15	63.84	6518.00	5561.00	346.63	119.99	70.71
16	80.47	6258.00	5379.00	354.98	122.57	70.92
17	76.70	5001.00	4091.00	318.63	118.17	70.35
18	59.95	3490.00	3280.00	273.56	116.20	68.83
19	61.33	2516.00	2059.00	270.49	108.91	71.66
20	71.76	411.00	222.00	86.91	107.26	92.24
Average=	69.34	5439.65	4739.53	322.77	121.40	72.58
	Tumor Volume (cm3):	**924.74**	**805.72**			

Table 3: Scan Series ID 2 statistics - 1

Here, we compare the results of the automatic detection of tumor regions to a manual detection by eye.

Summary

Medical instruments produce very valuable images.

Automatic detection of problematic regions (Wilms tumor) and classification of problematic cases (lung cancer – deep learning) are exciting developments.

In the future, there is hope to combine image analysis with e.g. simultaneous spectroscopic measurements.

Image detection will likely become a more central part of bioinformatics in future years.