TU/e Technische Universiteit
Eindhoven
University of Technology

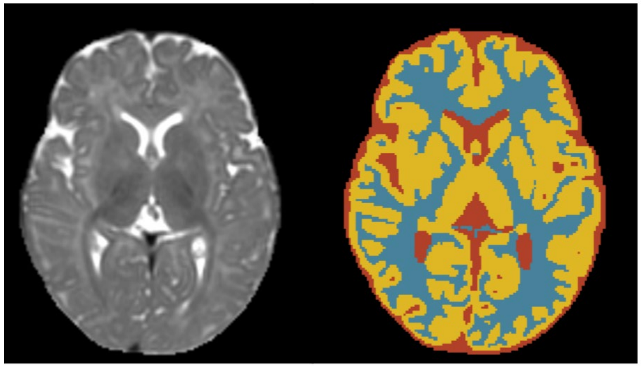
Validation in medical image analysis

& Active shape models

Maureen van Eijnatten

Overview of different medical image analysis tasks (2D, 3D, 3D+, …)

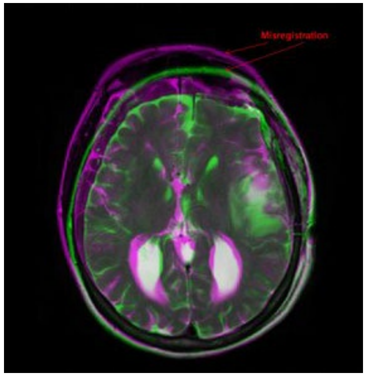
Image segmentation



Dividing an image into multiple regions with similar properties (e.g., intensity values).

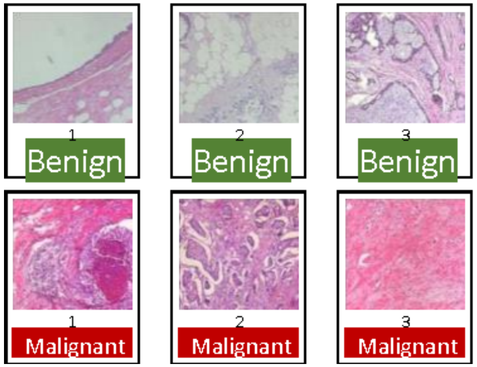
NB: these regions typically correspond to different anatomical structures.

Image registration



Finding an optimal transformation that aligns two images.

Computer-aided detection (CAD)



Categorizing/labeling images based on specific rules.

Official definition: “systems that assist doctors in the interpretation of medical images, often based on machine learning”

Outline for today

- **Validation in medical image analysis**
  - Image segmentation
  - Image registration
  - CAD

- **Active shape models** (= an image segmentation strategy)

Quality measures for medical image analysis

Task	Quality measure
Segmentation	Correspondence between the segmented object and a reference segmentation
Registration	Deviation from the correct transformation (e.g., TRE)
Computer-aided detection (CAD)	Ratio between correct and incorrect decisions

Recommended reading:

Chapter 13.1 of the Guide to Medical Image Analysis by Tonnies, Klaus D

Important characteristics to consider when evaluating medical image analysis methods:

- **Accuracy** = deviation of results from known ground truth.
- **Precision**, **reproducibility, reliability** = extent to which equal or similar input produces equal or similar results.
- **Robustness** characterizes the change of analysis quality if conditions deviate from assumptions made for analysis (e.g., when noise level increases or if object appearance deviates from prior assumptions).
- **Efficiency** = effort necessary to achieve an analysis result.

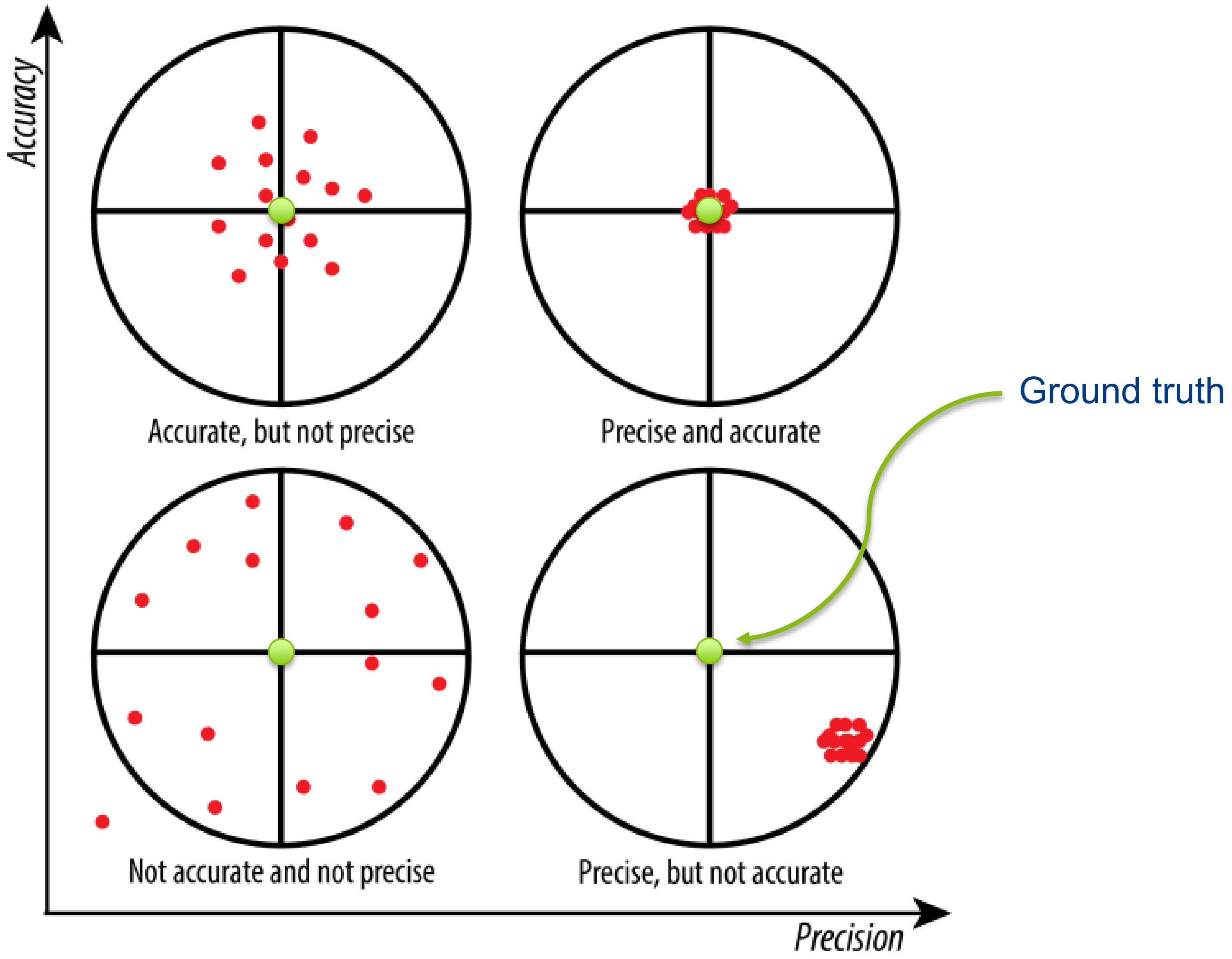

Accuracy
Accurate, but not precise
Precise and accurate
Ground truth
Not accurate and not precise
Precise, but not accurate
Precision

Image source: https://wp.stolaf.edu/it/gis-precision-accuracy/

ground truth = a conceptual term relative to the knowledge of the **truth** concerning a specific question (the "ideal expected result")

But the goal of medical imaging itself poses an inherent challenge…

"In medical image analysis, the truth is difficult to come by, since the reason for producing images in the first place was to gather information about the human body that cannot be accessed otherwise."

So, how can we get a **ground truth**?

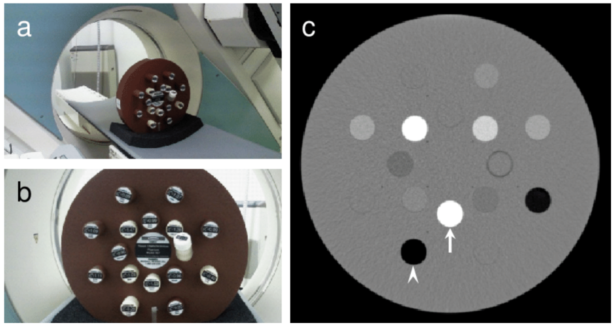
A. Based on *real data*

- Artificial hardware (imaging) phantoms
- Cadaveric material

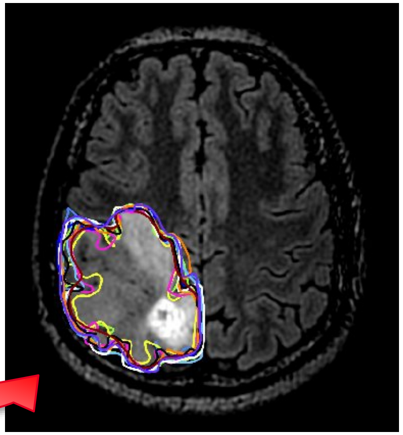
- Other imaging modality
- Other analysis method
- Expert annotations
  - e.g., radiologists, pathologists

(intra- & inter-observer variability?)

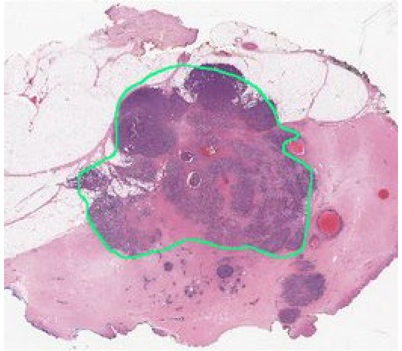
Anthropomorphic head phantom (CT)



Attenuation phantom (CT)



Manual glioma segmentation (MRI)

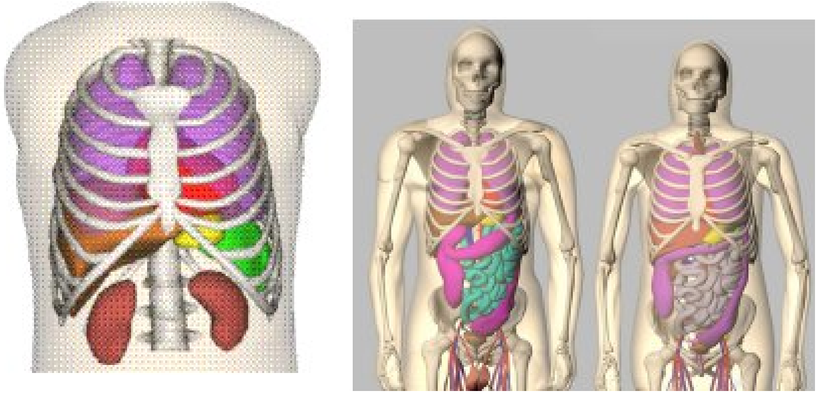


Tissue segmentation by pathologist

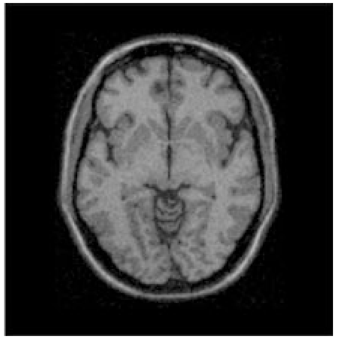
So, how can we get a **ground truth**?

B. Based on *simulated data*

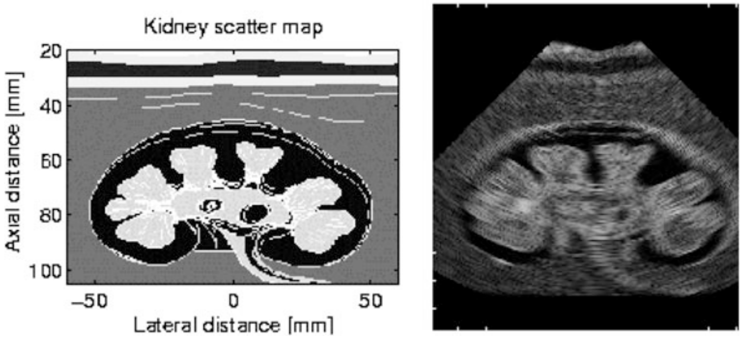
- Software phantom
  - E.g., XCAT phantom for PET validation, BrainWeb phantom, ultrasound phantom
- Mathematical simulations
  - E.g., Shepp-logan phantom



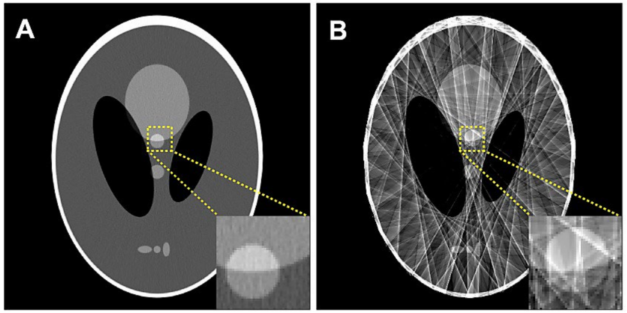
4D XCAT phantom



BrainWeb phantom (MRI)



Ultrasound phantom (Jensen and Svendsen 1992, 1996)



Shepp-logan (CT)

Validation of image *segmentation* methods

Example: evaluation of a segmentation task

- Which approach do we choose?

- Compare to ground truth:

  score = evaluate_segmentation(segmentation, ground_truth)

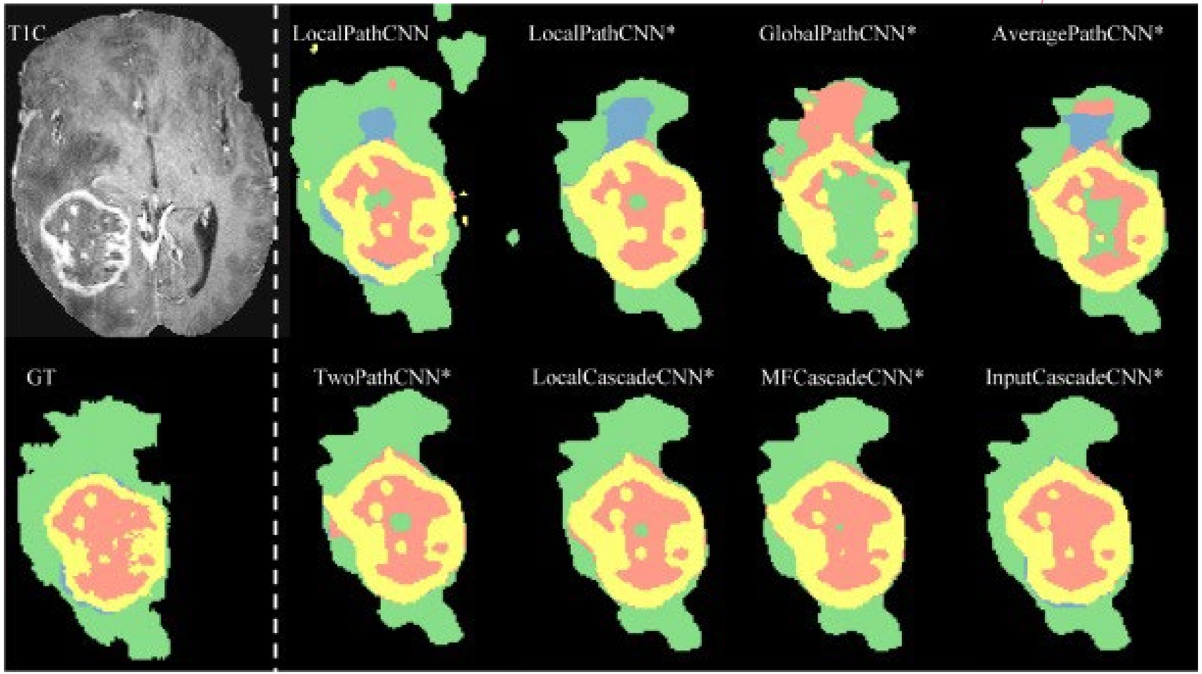


Image https://www.sciencedirect.com/science/article/pii/S1361841516300330

Evaluation metrics

Many metrics available, we look at:

- Accuracy
- Dice score
- Hausdorff distance

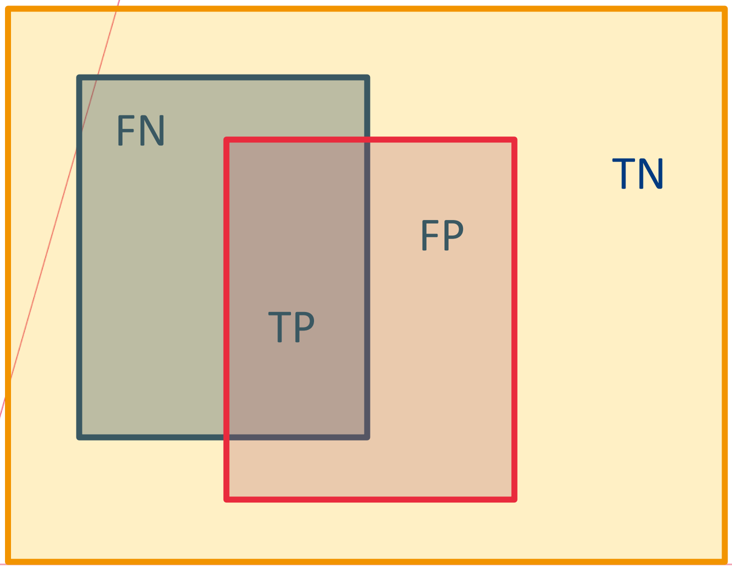
Accuracy

“How many pixels are correct?”

Accuracy = (TP + TN) / (TP+FP+FN+TN)

- TP = True Positive
- FP = False Positive
- FN = False Negative
- TN = True Negative

Orange = whole image
Blue = ground truth
Red= segmentation result



What if the ground truth is small?

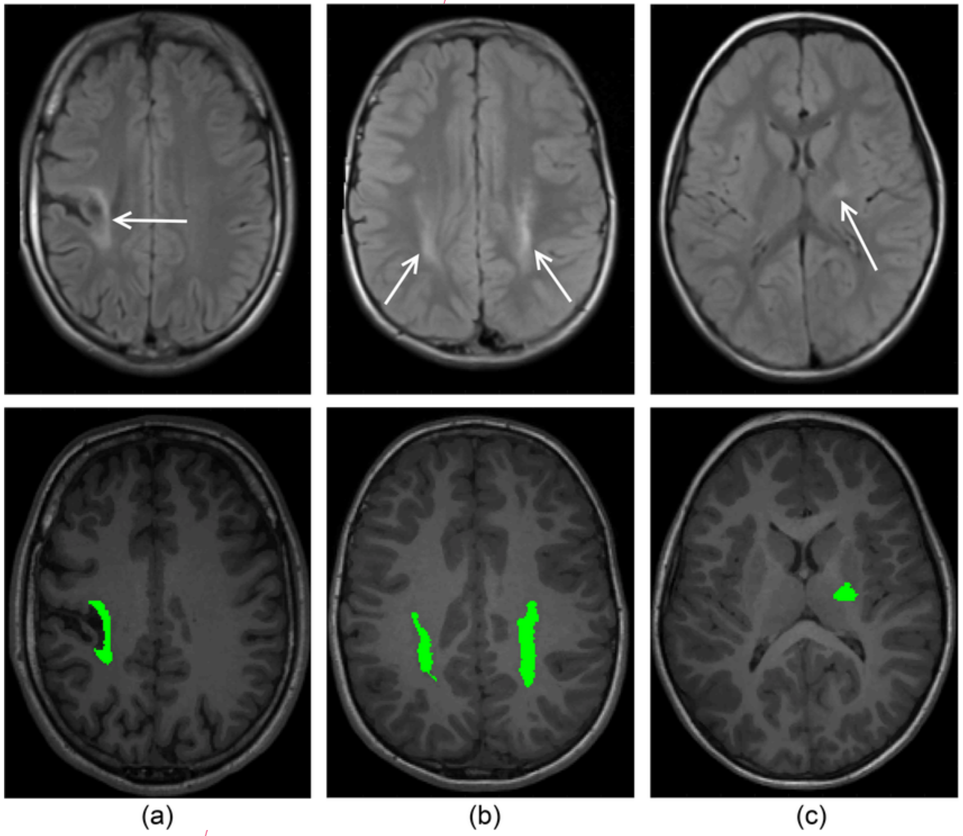


Image CC-BY 4.0 https://www.researchgate.net/figure/Illustration-of-lesion-segmentation-A-GM-lesion-segmentation-B-WM-lesion_fig4_318992008

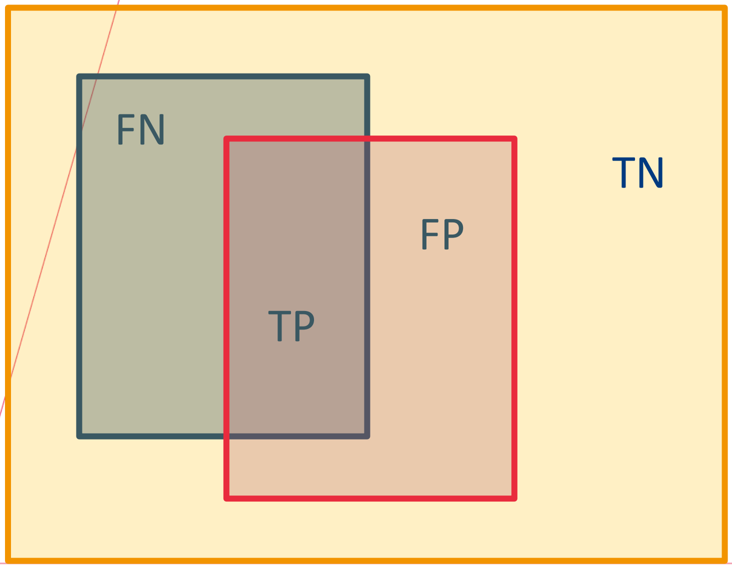
Dice score

Measure overlap excluding TN

Sørensen–Dice index a.k.a. Dice Similarity Coefficient (DSC)

DSC = 2TP / (2TP + FP + FN)

Orange = whole image
Blue = ground truth
Red= segmentation result



Dice score

Two equivalent definitions

DSC = 2TP / (2TP + FP + FN)

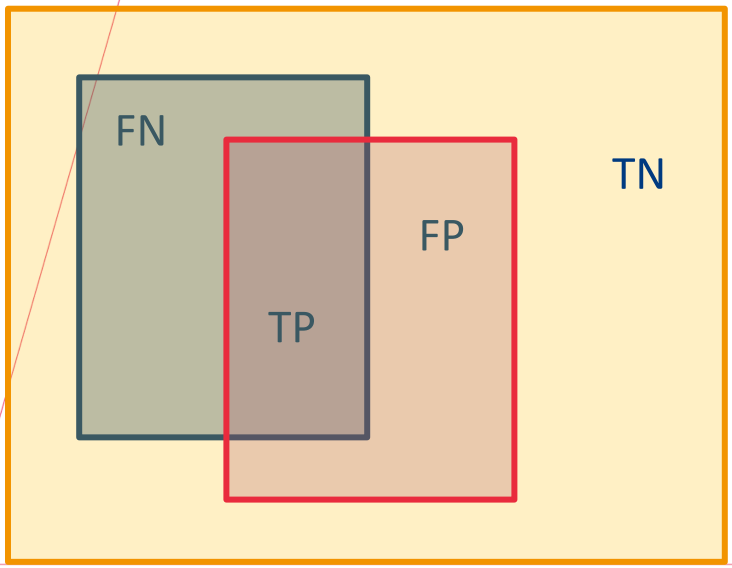
DSC = 2 | A * B | / (| A | + | B |) for binary images A and B

| A | = size of blue ground truth

| B | = size of red result

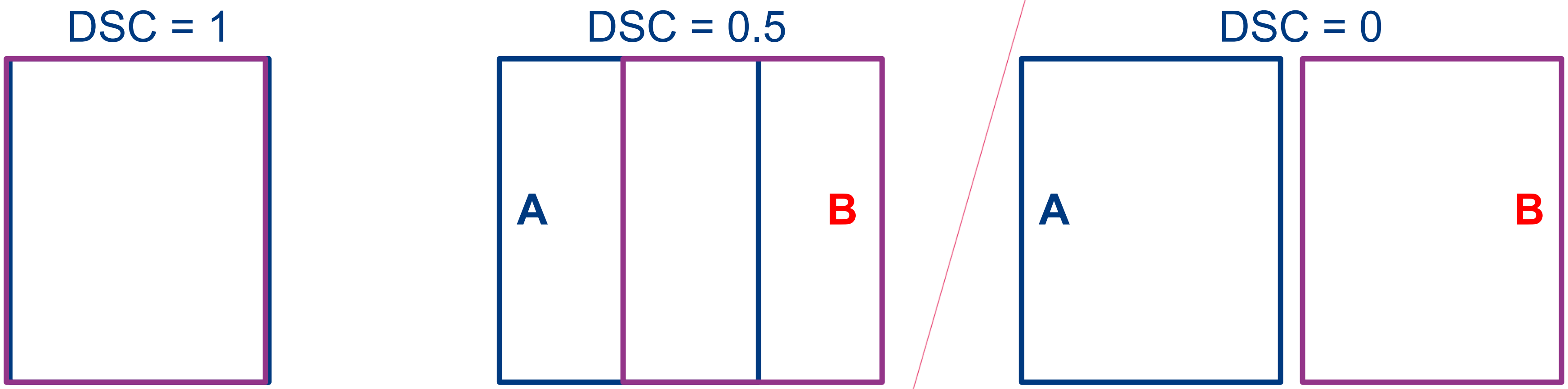
| A * B| = size of overlap

Orange = whole image
Blue = ground truth
Red= segmentation result



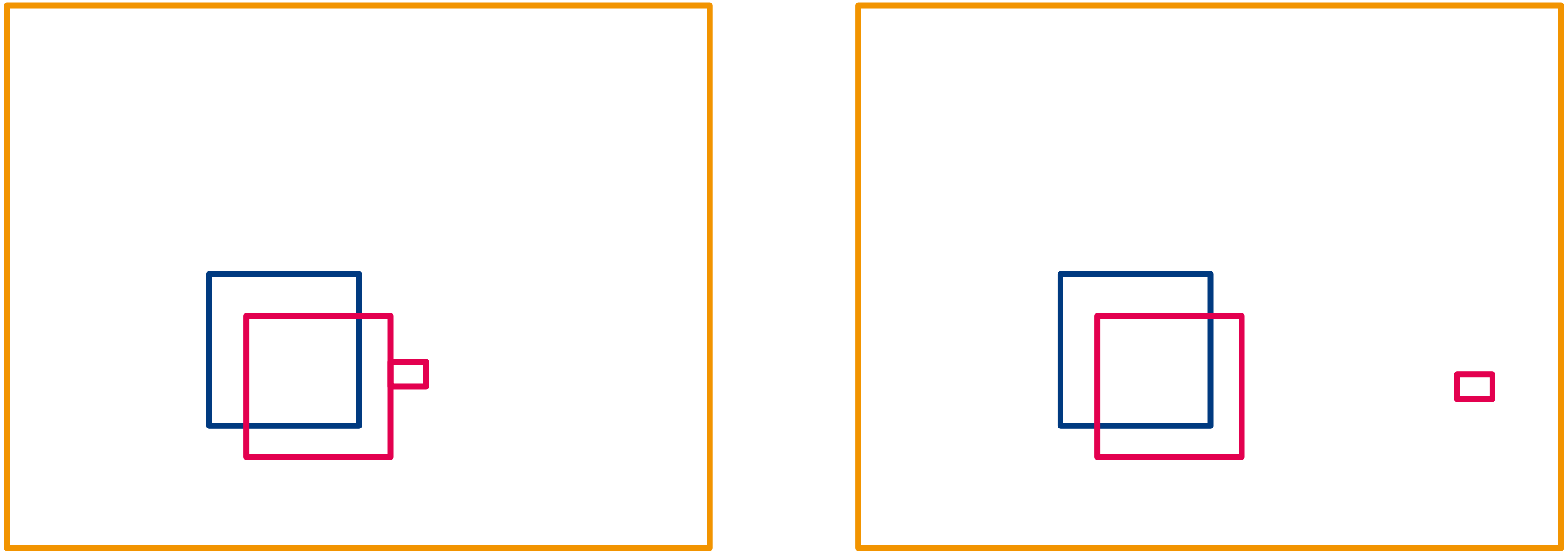
Dice score

- Between 1 and 0 for full / no overlap



Limitations of the Dice score

Dice (and other metrics based on TP, FP, etc.) are not sensitive to location



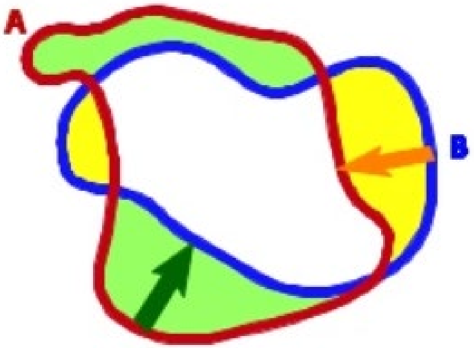
Hausdorff distance

Compare sets of points on the boundaries

Hausdorff distance = maximum shortest distance between the boundary points

$$\mathrm{h}(\mathrm{A},\mathrm{B}) = \max_{a\,\in A}\min_{b\,\in B} d(a,b)$$

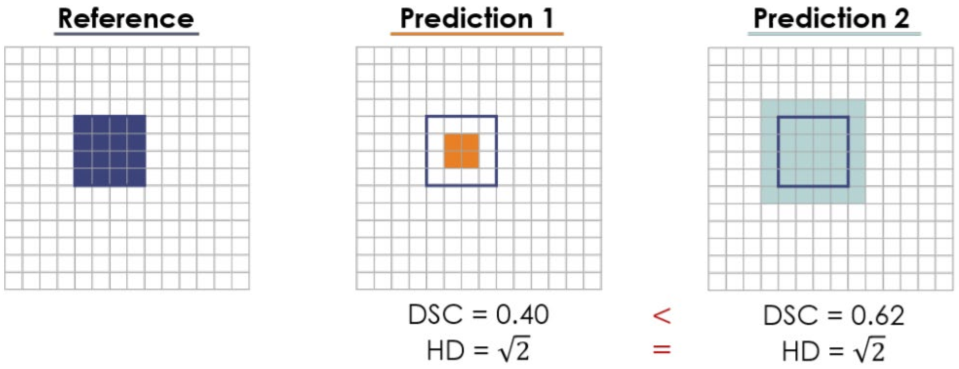
$$H(\mathrm{A},\mathrm{B}) = \max(h(A,B),h(B,A))$$



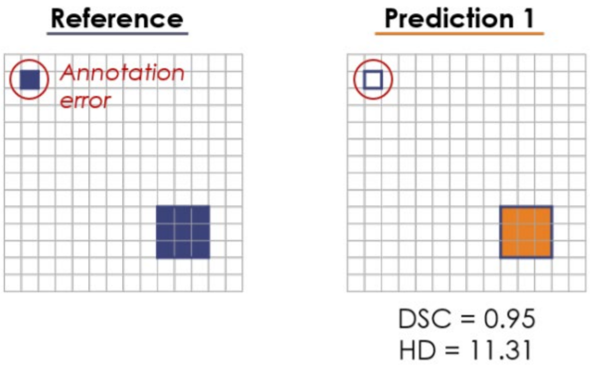
https://www.slideshare.net/UlaBac/lec14-evaluation-framework-for-medical-image-segmentation

Hausdorff distance (HD) takes location into account and represents over- and undersegmentation equally, but is also more sensitive to outliers/errors:

Over/undersegmentation:



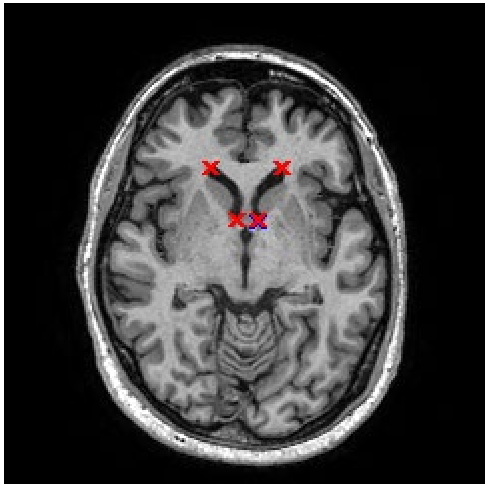
Outliers:



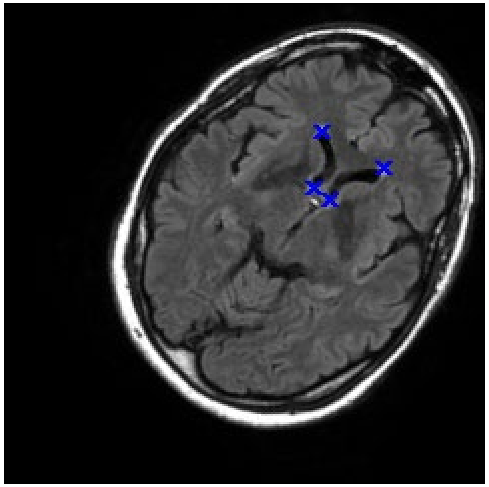
From: Common Limitations of Image Processing Metrics: A Picture Story (https://arxiv.org/pdf/2104.05642.pdf)

Validation of image *registration* methods

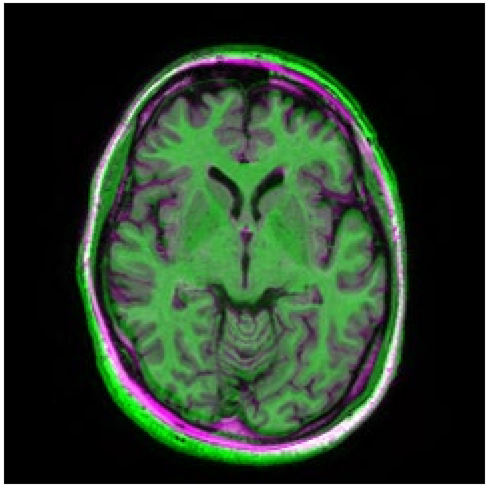
Remember this example:



Fixed



Moving

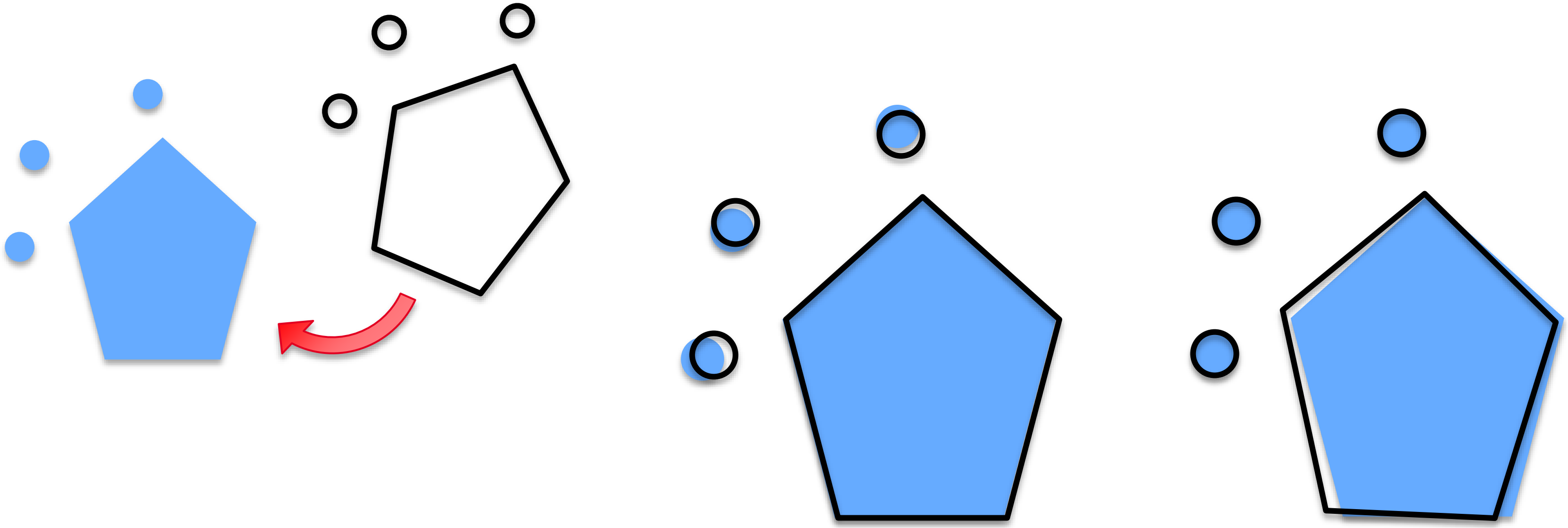


Registered

Measure registration performance by computing the registration error for some target point pairs. These target points must be different from the points used to compute the transformation! Why?

The **optimal fit** for the fiducial markers does not automatically mean that the registration itself is optimal, especially if:

- the markers are far away from the object to be registered
- too few markers are used
- it is difficult to localize the markers.

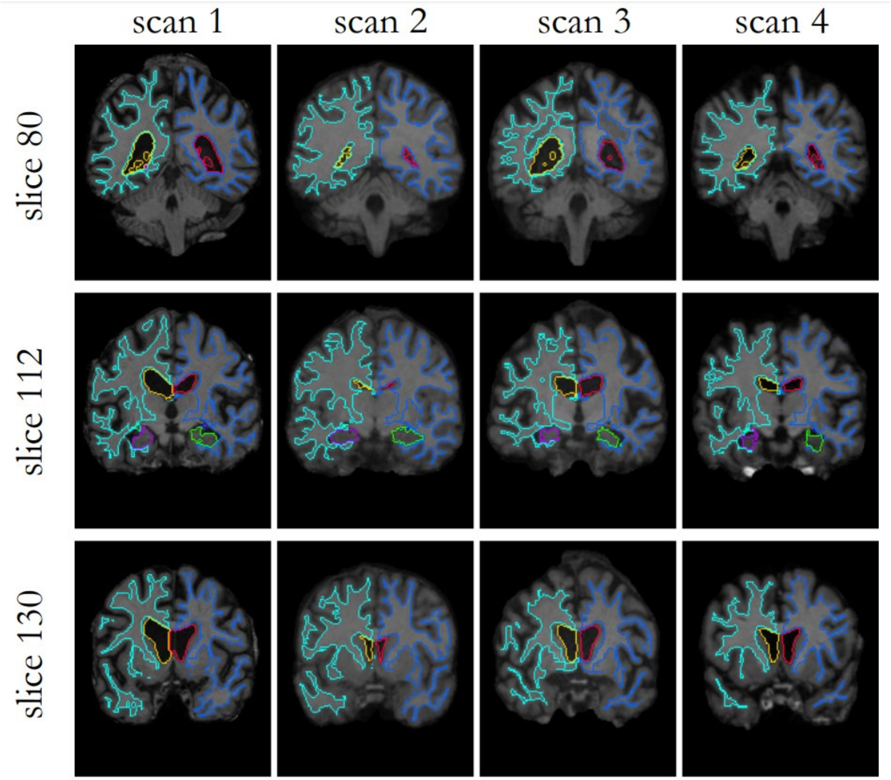


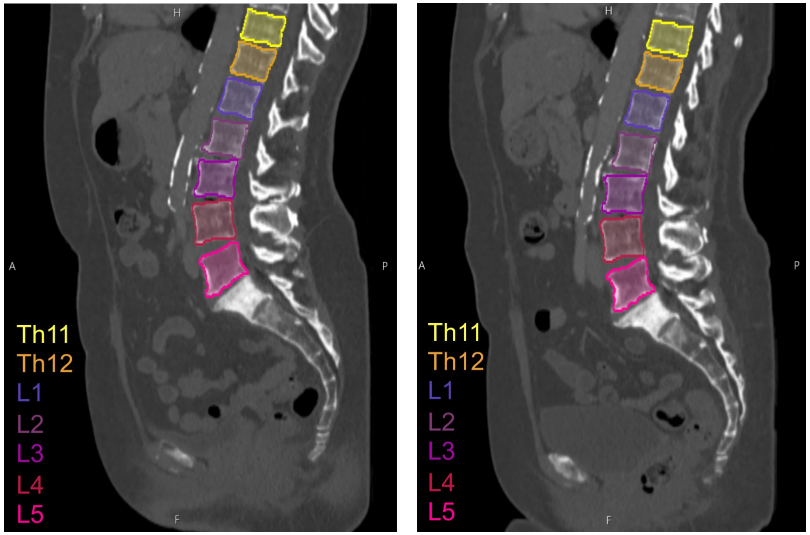
Optimal
registration

Optimal fit
fiducial markers

How can we measure the quality of a **registration** task?

Idea: apply the transformation to a *segmentation* mask that represents the anatomy of interest.





Balakrishnan, Guha, et al. "Voxelmorph: a learning framework for deformable medical image registration." *IEEE transactions on medical imaging* 38.8 (2019): 1788-1800.

Validation of *CAD* methods

Dice score is ***not*** suitable for classification/detection tasks

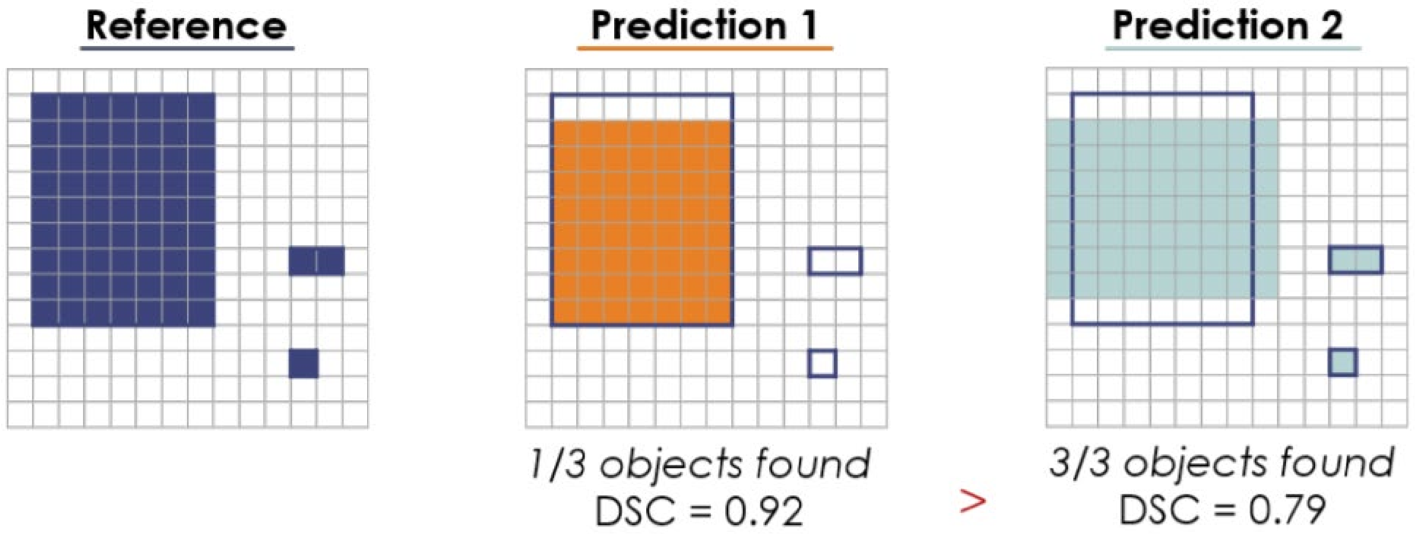


Figure 6 Effect of using a segmentation metric for object detection. In this example, the prediction of one algorithm only detecting one of three structures (*Prediction 1*) leads to a higher *DSC* compared to that of a second algorithm (*Prediction 2*) detecting all structures.

From: Common Limitations of Image Processing Metrics: A Picture Story (https://arxiv.org/pdf/2104.05642.pdf)

Relevant evaluation metrics for CAD based on TP, FP, etc. are the **sensitivity** and **specificity.**

- Sensitivity = TP / TP + FN
- Specificity = TN / TN + FP

		Reality / Ground truth	
		✓	✗
Measured	✓	True Positives (TP)	False Positives (FP) “Type I error”
	✗	False Negatives (FN) “Type II error”	True Negatives (TN)

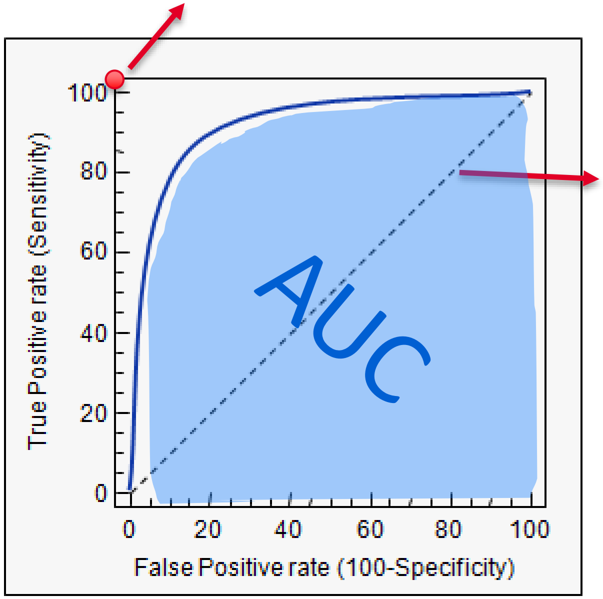
NB: These metrics are commonly used in **detection** tasks involving medical images. Interestingly, they are also very important when interpreting the performance of any test (e.g., airport security, breast cancer screening, quality assurance in companies, …)

There is usually a trade-off between sensitivity and specificity:

Receiver operator characteristic (ROC) curve

& Area Under the Curve (AUC)

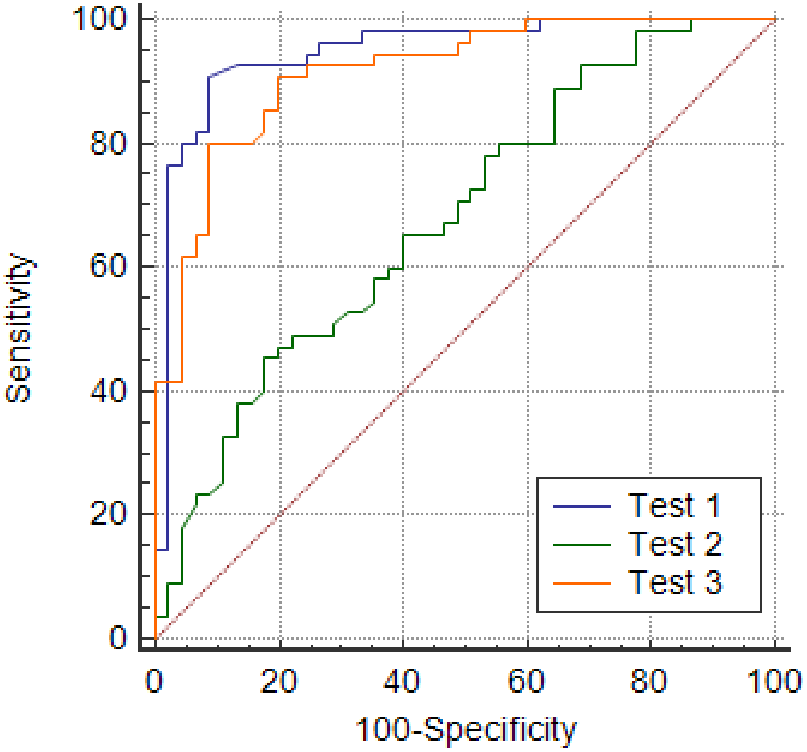
Perfect classifier

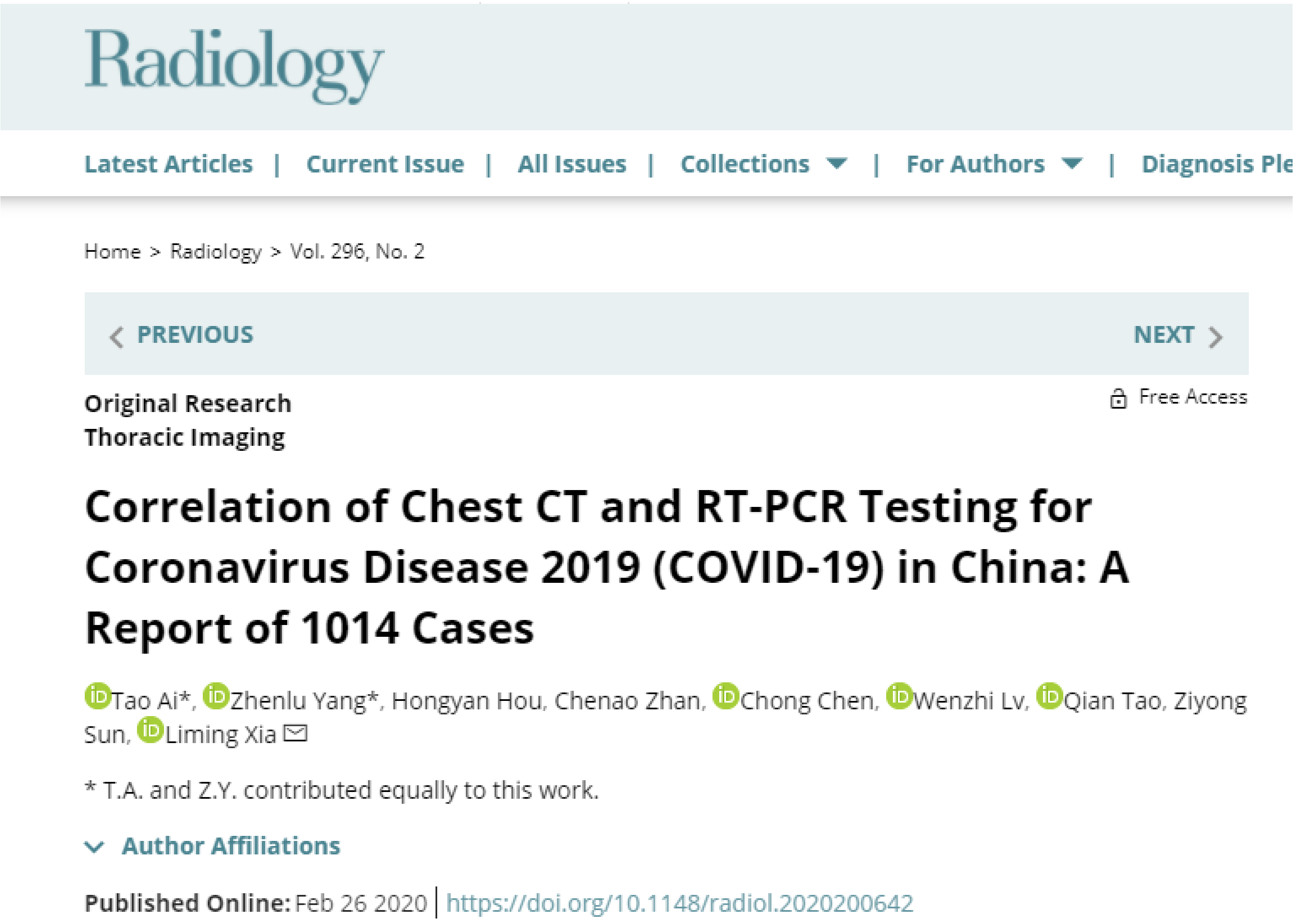


No diagnostic value

Receiver operator characteristic (ROC) curve & Area Under Curve (AUC)

Question: What is the best test?




Radiology
Latest Articles | Current Issue | All Issues | Collections | For Authors | Diagnosis Ple
Home > Radiology > Vol. 296, No. 2
PREVIOUS
NEXT
Free Access
Author Affiliations

Original Research
Thoracic Imaging

Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases

Tao Ai*, Zhenlu Yang*, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, Liming Xia

* T.A. and Z.Y. contributed equally to this work.

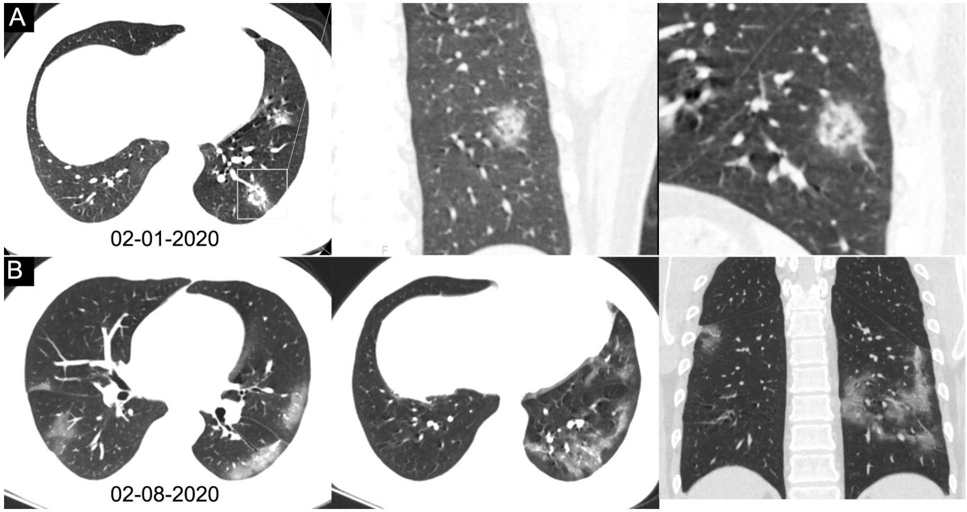
Published Online: Feb 26 2020 | https://doi.org/10.1148/radiol.2020200642

Ai T, Yang Z, Hou H, et al. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology*. 2020;296(2):E32-E40.

Results

“Of the 1014 patients, 601 of 1014 (59%) had positive RT-PCR results and 888 of 1014 (88%) had positive chest CT scans. The sensitivity of chest CT in suggesting COVID-19 was 97% (95% confidence interval: 95%, 98%; 580 of 601 patients) based on positive RT-PCR results.

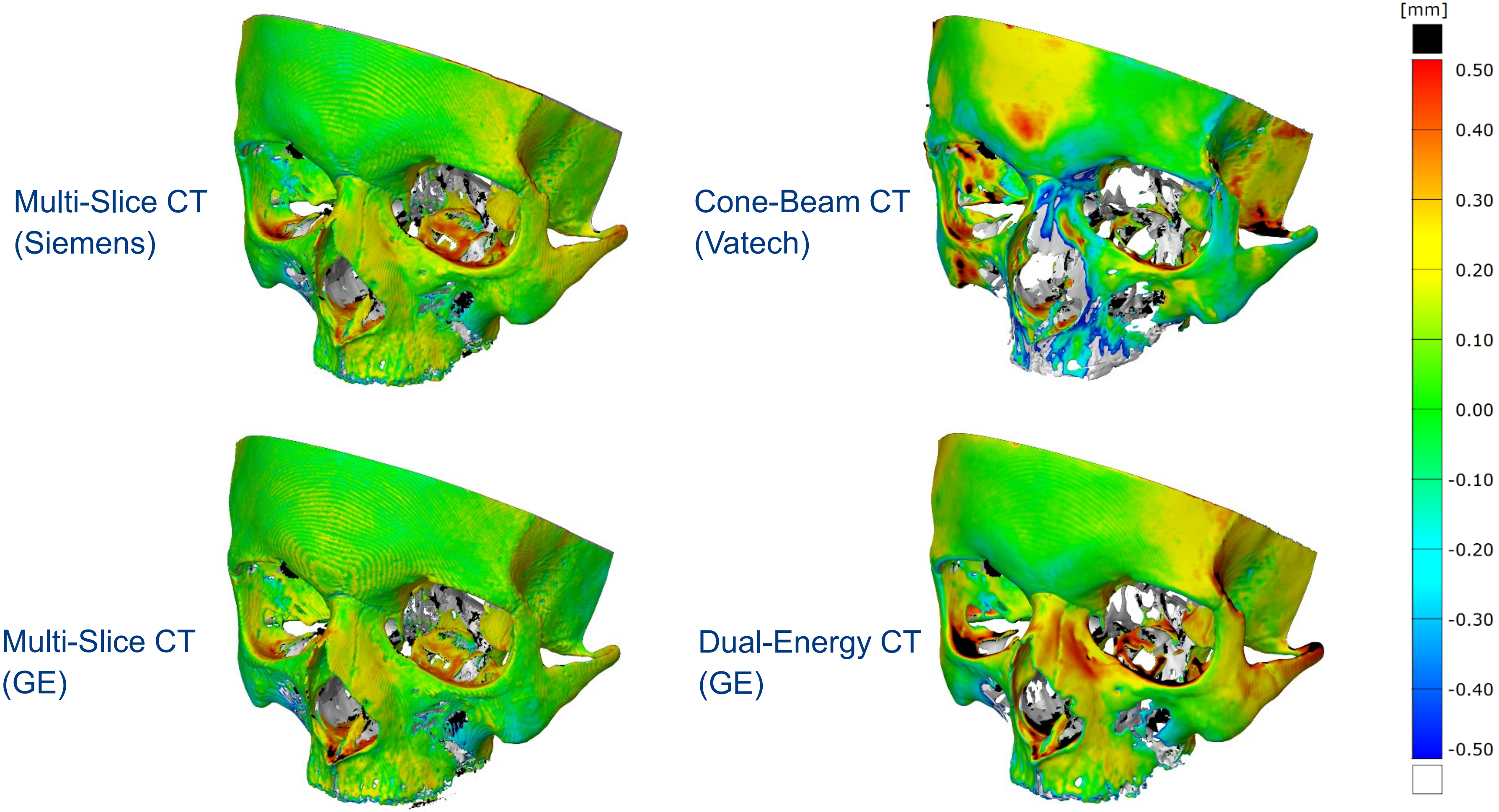
In patients with negative RT-PCR results, 75% (308/413) had positive chest CT findings; of 308, 48% were considered as highly likely cases, with 33% as probable cases.”



What about the specificity?

Ai T, Yang Z, Hou H, et al. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology*. 2020;296(2):E32-E40.

Measuring and/or visualizing surface distances can give fascinating insights into the accuracy of a medical image analysis task:



Multi-Slice CT (Siemens)

Cone-Beam CT (Vatech)

Multi-Slice CT (GE)

Dual-Energy CT (GE)

van Eijnatten, M., Berger, F. H., de Graaf, P., Koivisto, J., Forouzanfar, T., & Wolff, J. (2017). Influence of CT parameters on STL model accuracy. Rapid Prototyping Journal, 23(4), 678-685.

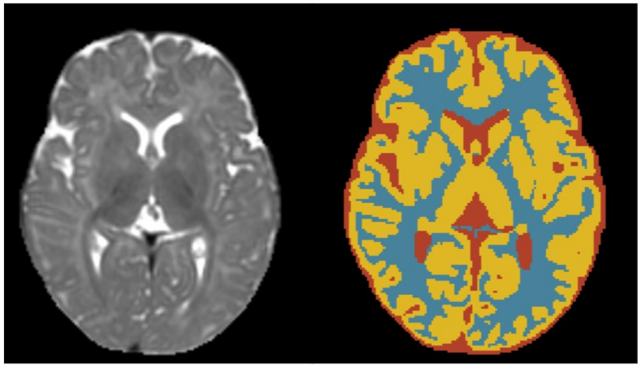
Further reading:

- Guide to Medical Image Analysis - Methods and Algorithms
  https://link.springer.com/book/10.1007/978-1-4471-2751-2

Active shape models

Overview of different medical image analysis tasks (2D, 3D, 3D+, …)

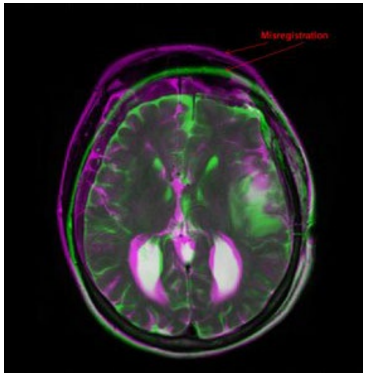
Image segmentation



Dividing an image into multiple regions with similar properties (e.g., intensity values).

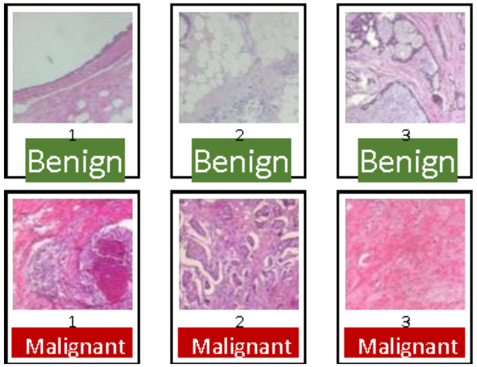
NB: these regions typically correspond to different anatomical structures.

Image registration



Finding an optimal transformation that aligns two images.

Computer-aided detection (CAD)



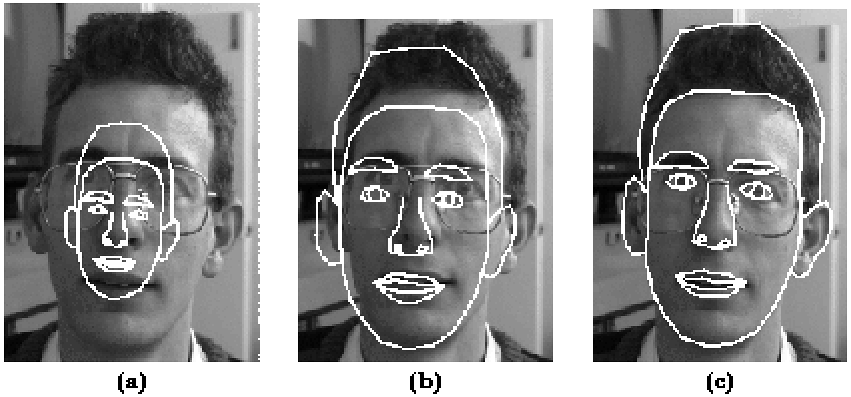
Categorizing/labeling images based on specific rules.

Official definition: "systems that assist doctors in the interpretation of medical images, often based on machine learning"

Shape models

Motivation: segmenting individual structures in (medical) images

We can expect structures to have a certain shape – how do we use this information?



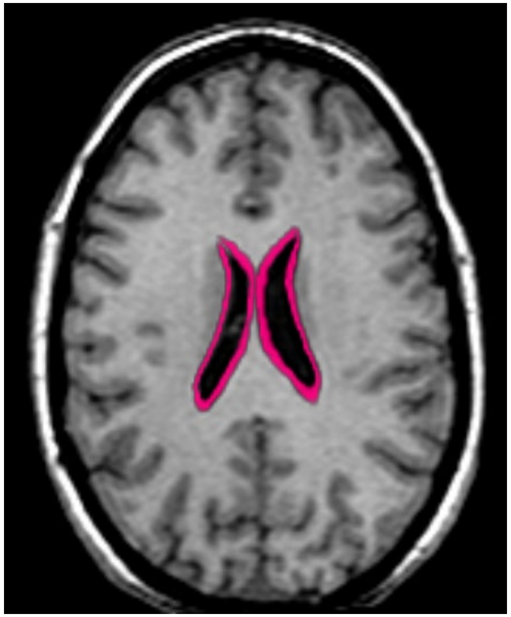


Image BIGR / Stefan Klein

Shape models

- Model shape using
  - prior knowledge (e.g. the shape will be round)
  - “typical shape” training data
- Fit model to test image as well as possible
- Many methods in this family, e.g.
  - Snakes (8DBOO)
  - Deformable templates
  - Active shape models

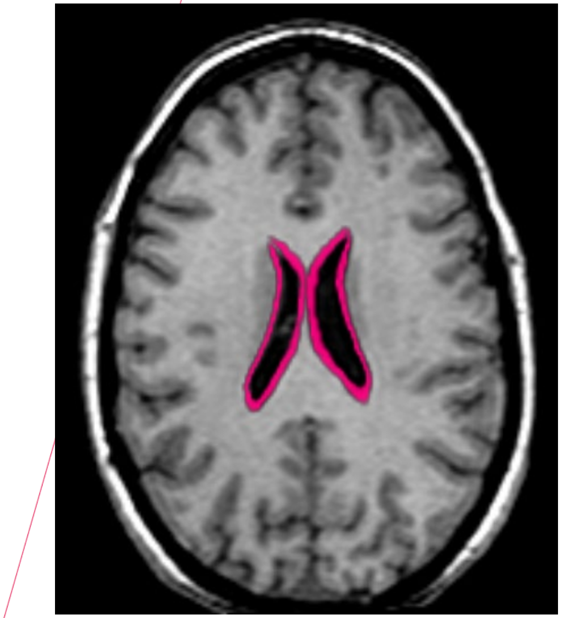
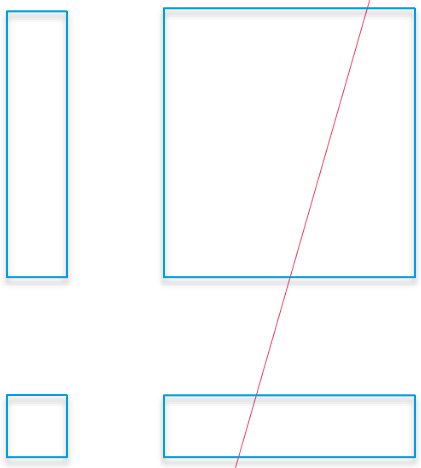


Image BIGR / Stefan Klein

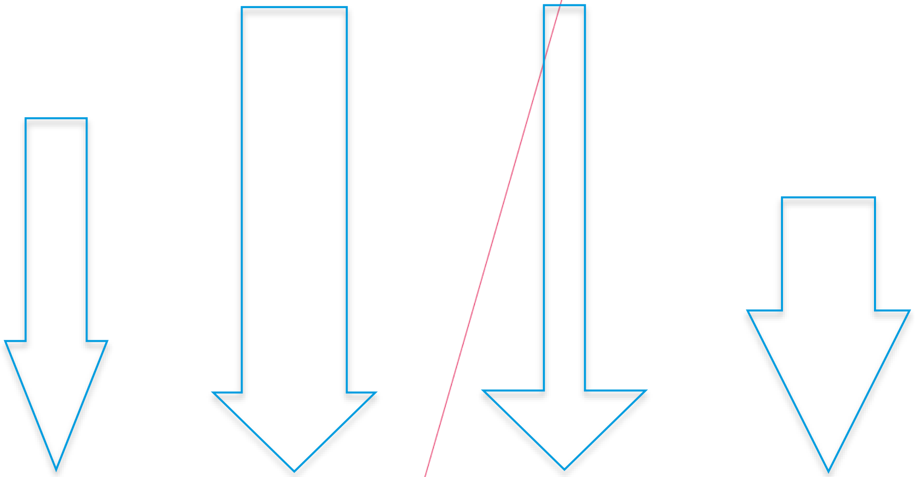
Shape model example

Rectangles

Two parameters: width, height



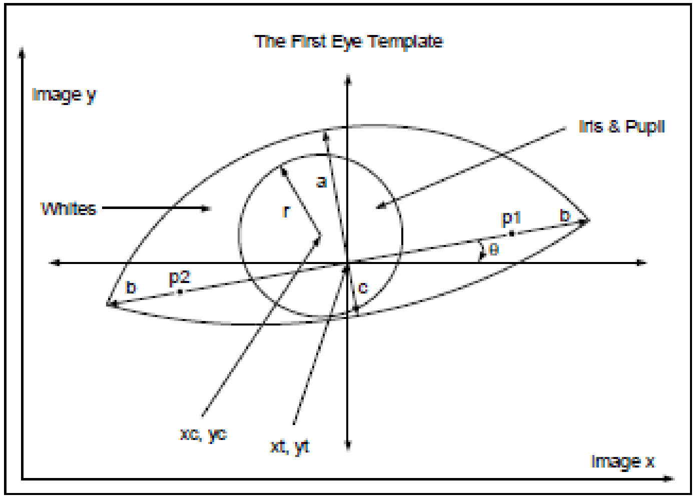
Shape model example



How many parameters?

Example

Parameters: various lengths and angles



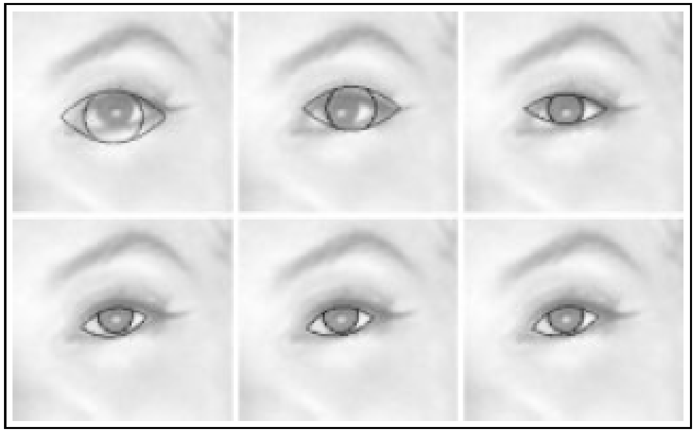


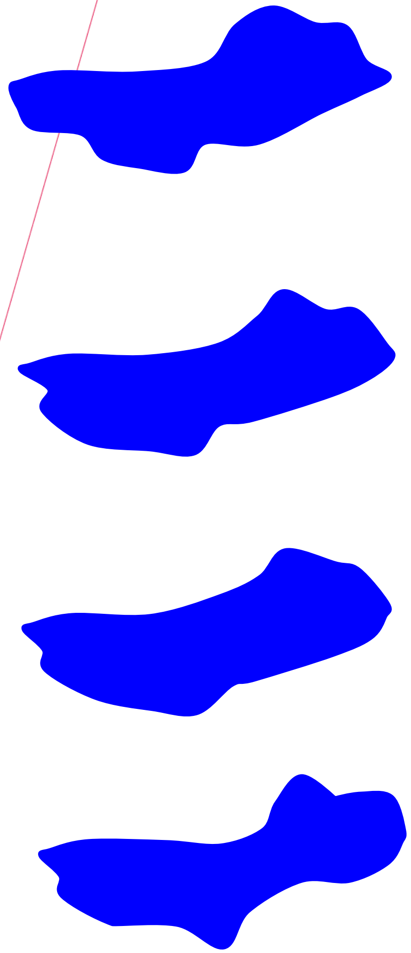
Figure 2.5: Deformable eye template *An eye template is defined (top) in terms of a modest number of variable geometric parameters. In successive iterations of a "gradient descent" algorithm, an equilibrium configuration is reached in which the template fits the eye closely. (Figure reprinted from (Yuille and Hallinan, 1992) which also gives details of external and internal energy functions.)*

Image from Andrew Blake and Michael Isard. Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion. Springer, 1998.

Shape models

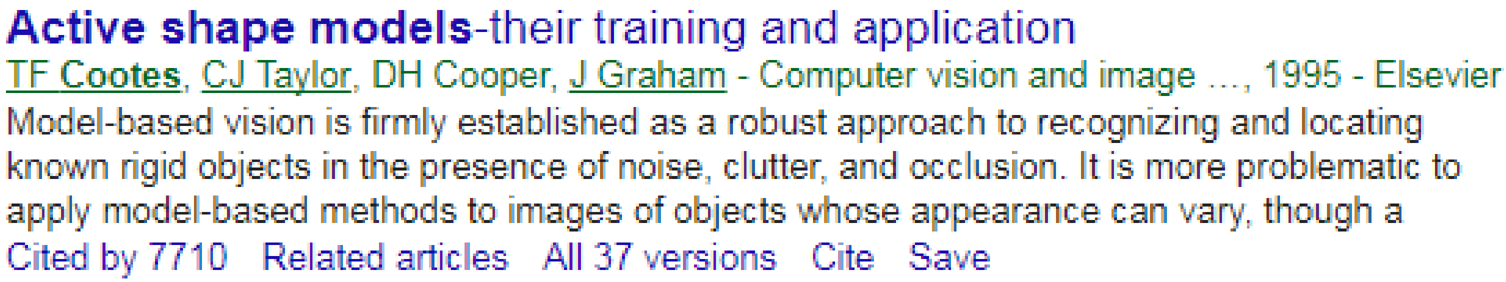
How can we model a shape that isn’t easily described by lines, circles etc?

- Cannot use parameters like “width” anymore
- Instead, place *K* points on the boundary
- In 2D, this leads to *2K* features (x,y coordinates)
- The shapes need to be aligned



Active shape models

- Paper / tutorial (**required reading**): https://www.sciencedirect.com/science/article/pii/S1077314285710041



Active shape models-their training and application

TF Cootes, CJ Taylor, DH Cooper, J Graham - Computer vision and image …, 1995 - Elsevier

Model-based vision is firmly established as a robust approach to recognizing and locating known rigid objects in the presence of noise, clutter, and occlusion. It is more problematic to apply model-based methods to images of objects whose appearance can vary, though a

Cited by 7710   Related articles   All 37 versions   Cite   Save

Image: Google Scholar

Active shape models: idea behind the model

- M-dimensional shape (usually M=2 or M=3)
- Represented by K boundary points in a *shape feature vector*

$$\mathbf{x} = (x_{1,1}\ x_{1,2}, \ldots, x_{1,K}\ x_{2,1}\ x_{2,2}, \ldots, x_{2,K}, \ldots, x_{M,1}, \ldots, x_{M,K}),$$

where $x_{m,k}$ is the m-th component of the k-th boundary point.

- Example data for active shape models

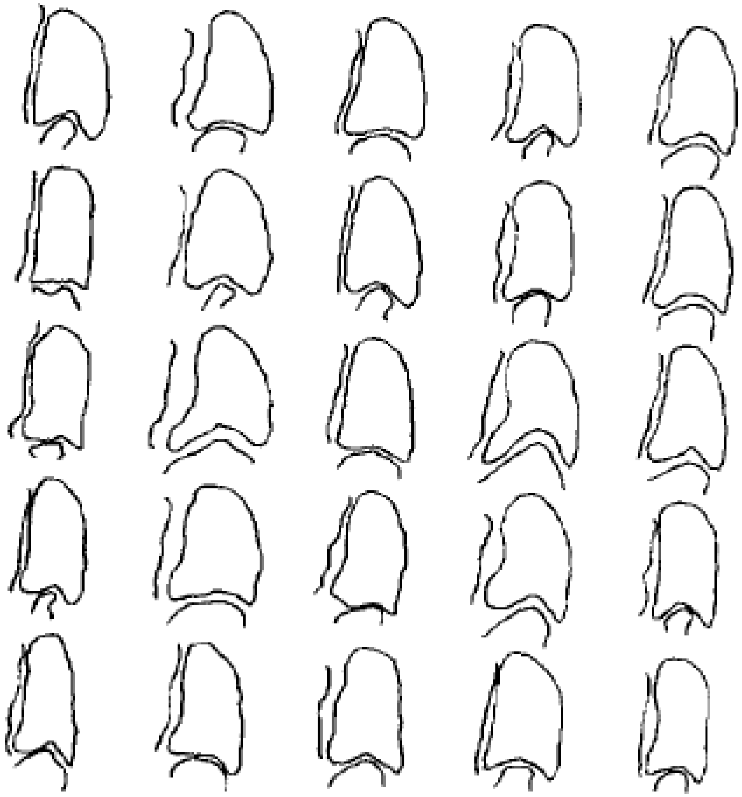
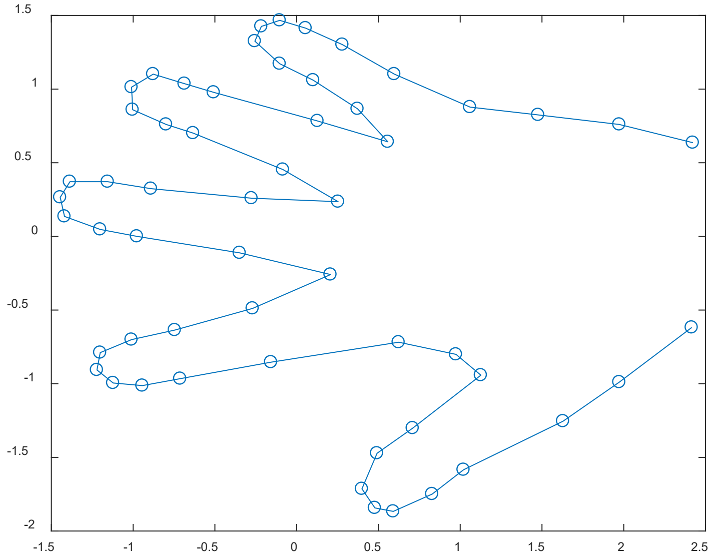
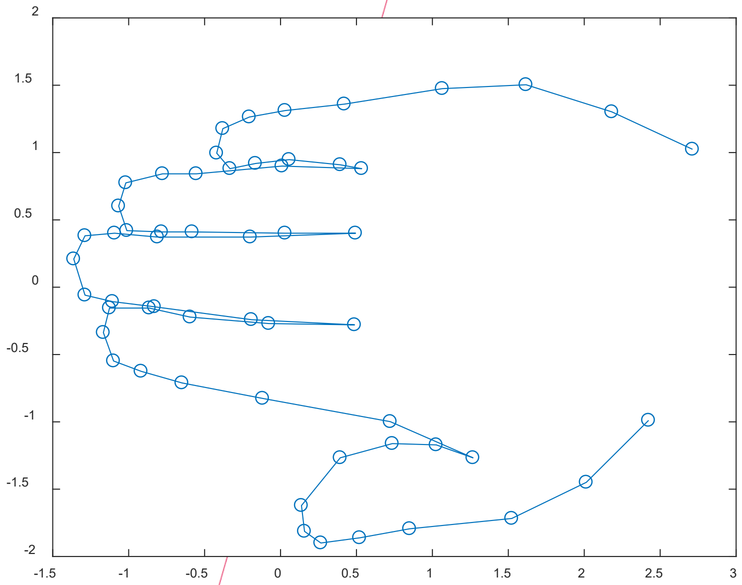


FIG. 10. Examples of heart ventricle shapes, each containing 96 points.

Image from Active shape models – their training and application, Cootes et al, 1995

- Example data for active shape models

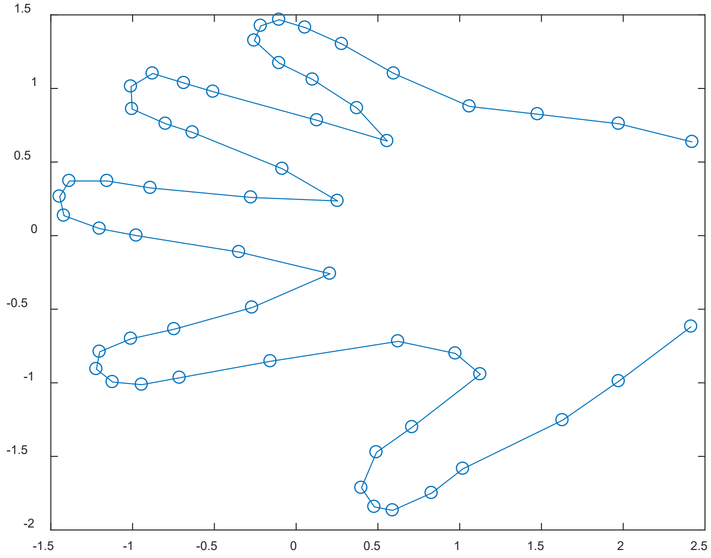


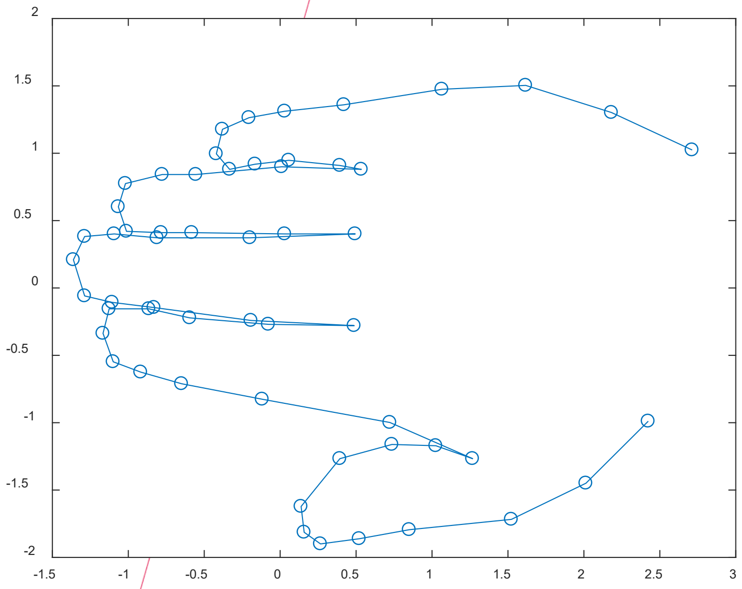


Active shape models: idea behind the model

- Are all shapes/variations probable?

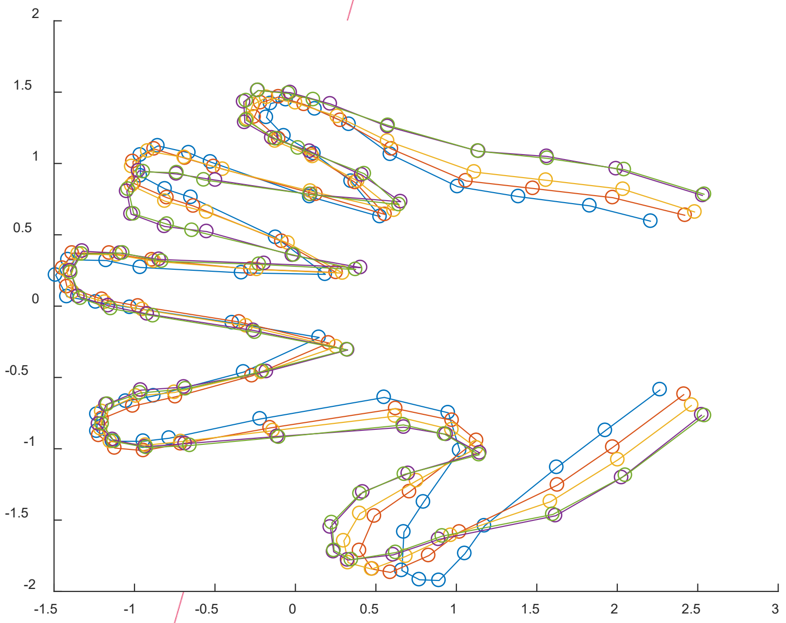
- Do we **need** 2K features to describe the variation?





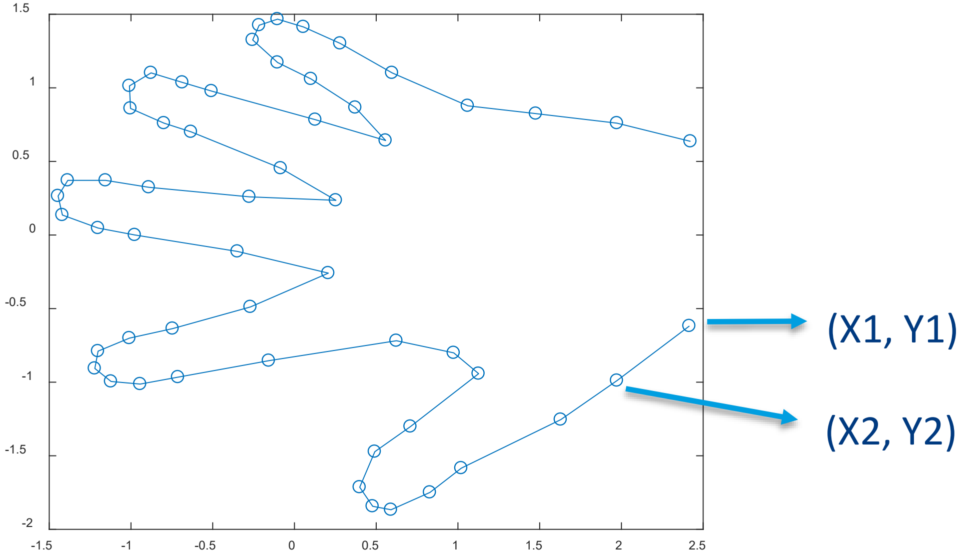
Active shape models: idea behind the model

- Not all shapes are probable, for example the points at the tips of the fingers will vary together
- "Length of fingers" is not a feature in our 2K dimensional space, but a combination of features
- We can store our model in less than 2K parameters



Active shape models: idea behind the model

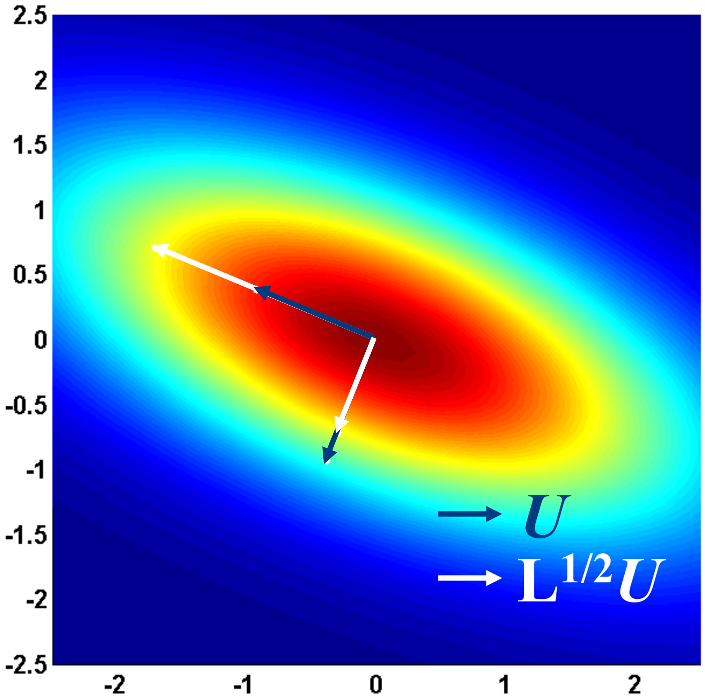
- Use Principal Component Analysis (PCA) to find main modes of variation



X1	X2	…	Y1	…
-0.1	0.2		1	
-0.2	0.1		1.1	
-0.1	0.2		0.9	
-0.2	0.2		1	

- Eigenvector = combination of existing features which represents a mode of variation

- Eigenvalue = how much variation is there



$$\Sigma = \begin{bmatrix} 3 & -1 \\ -1 & 1 \end{bmatrix} \longrightarrow \quad U = \begin{bmatrix} -0.92 & -0.38 \\ 0.38 & -0.92 \end{bmatrix} \quad \Lambda = \begin{bmatrix} 3.41 & 0 \\ 0 & 0.59 \end{bmatrix}$$

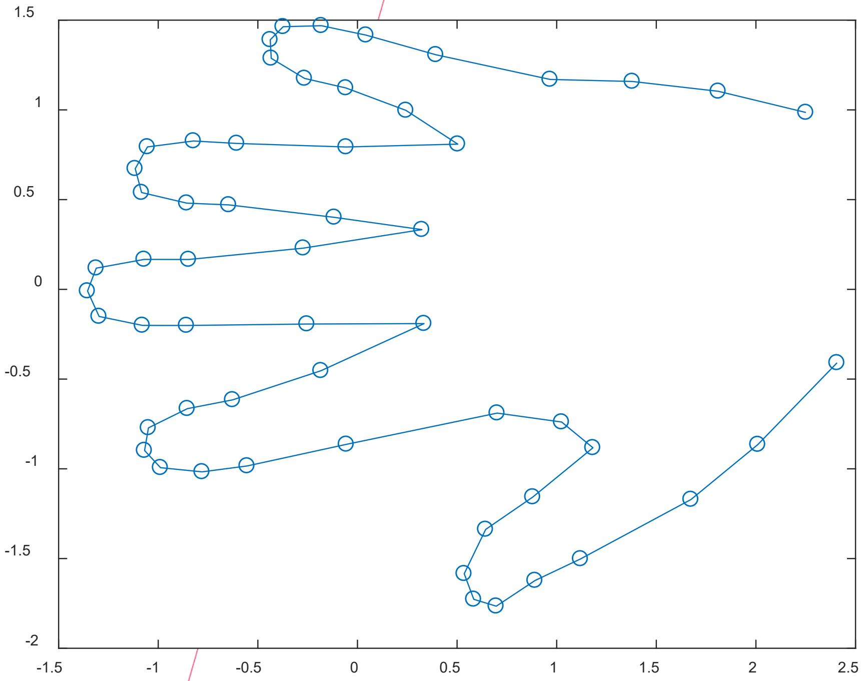
Steps for active shape model with 2K features

- Find the mean shape

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

- Find deviation from mean shape

$$d\mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$$



Steps for active shape model with 2K features

- Covariance matrix of the deviations (which coordinates often deviate together)

$$\sum \quad = \frac{1}{N}\sum_{i=1}^{N} d\mathbf{x}_i d\mathbf{x}_i^{\mathrm{T}}$$

- Eigendecomposition of covariance matrix → 2K eigenvectors / modes of variation & corresponding eigenvalues

$$\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_{2K}$$

- We do not need all 2K eigenvectors to describe most variation in the data. Can use fraction of variance to select only *f* eigenvectors (with highest eigenvalues)

Steps for active shape model with 2K features

We can see the *f* selected eigenvectors as a matrix $\mathbf{U}_f$

Any allowed shape can be approximated described as mean + linear combination of eigenvectors

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{U}_f\mathbf{b}$$

b is a vector of weights, each weight corresponds to how much variation we want along that eigenvector

Example: adding more or less variation along the first eigenvector

We change only 1 weight, but several of the original 2K features are affected

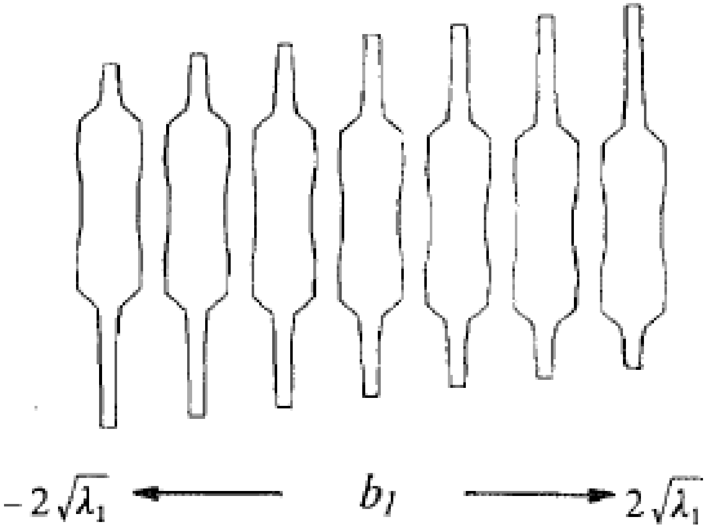


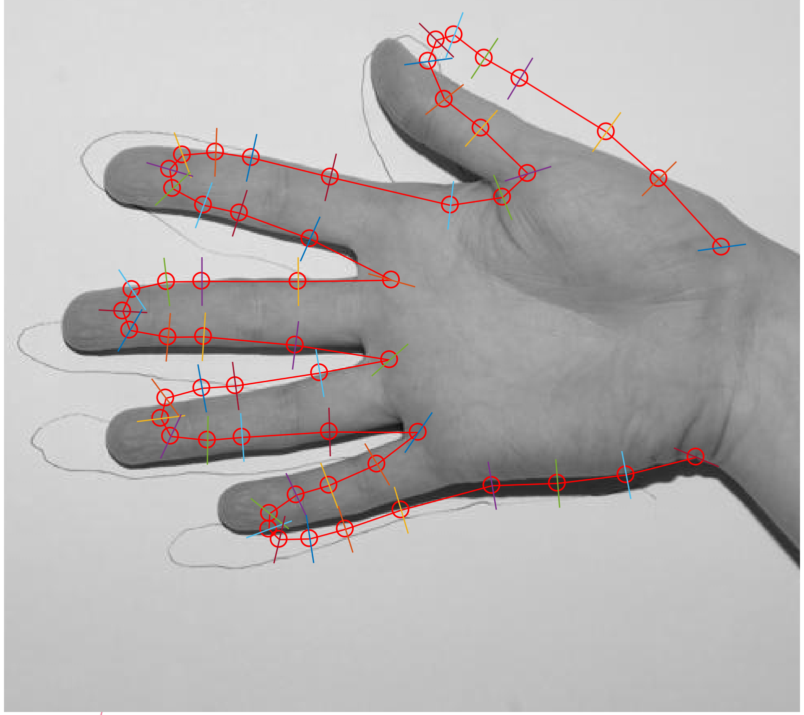
FIG. 7. Effects of varying the first parameter of the resistor model.

Image from Active shape models – their training and application, Cootes et al, 1995

Apply model to new image

Goal: Find 2K coordinates in a test image, such that these coordinates can be described by our shape model

Our model: weights **b** (shape), but also rotation/scaling of the shape (pose)

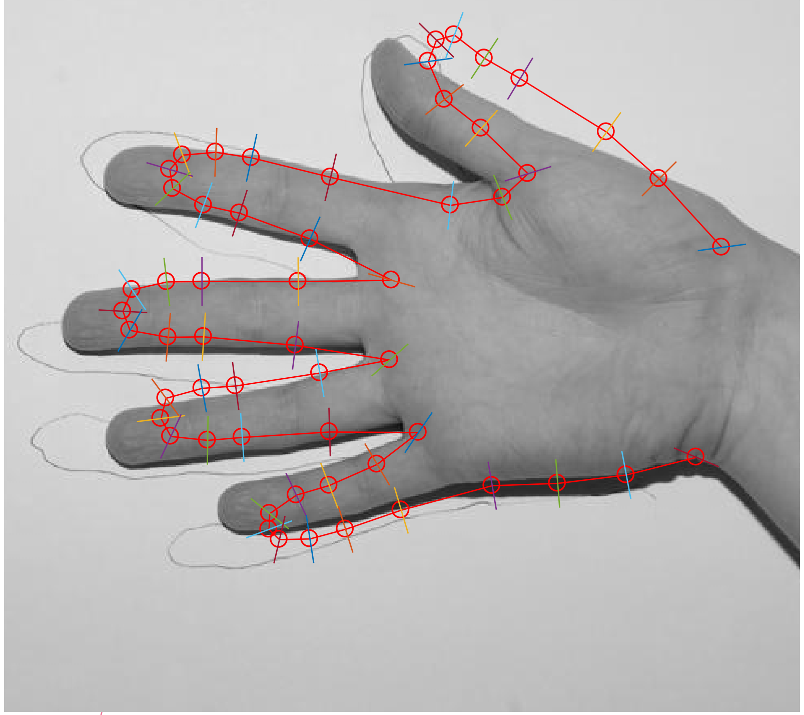


Apply model to new image

1) Start with initial position of points, X (2K vector)

2) Find translation vector dX that moves each point to a better position (close to an edge)

This is possible by looking at the intensity profile along the normal vector at each point

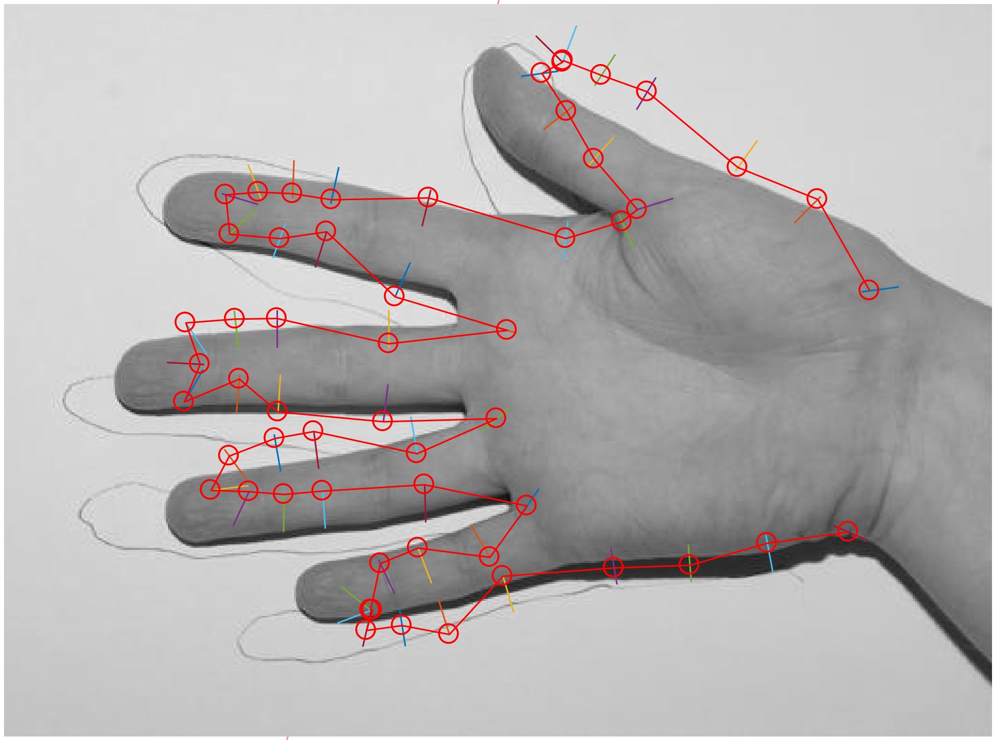


Apply model to new image

Just moving each point to a better position is not enough! X+dX is not a valid shape, so:

3) Find shape and pose, such that model(shape, pose) is close to X+dX

4) Repeat steps 2 and 3 until (almost) no change in dX



Active shape models - Generalization

- Training shapes must be representative for future data
- Too little variation in shapes → underfitting
- Too much variation → overfitting (can fit any shape)

You should be able to

- For a simple dataset of shapes, explain how many parameters are needed to model the variation
- Describe the steps needed to train an active shape model, and to apply it to a new image
- Motivate whether an active shape model is suitable for a particular dataset
- Reason about which coordinates in an active shape model might have a lot / little variation

Further reading

Guide to medical image analysis: https://link.springer.com/book/10.1007/978-1-4471-2751-2