Image Filtering and Image Segmentation: An Insight of Evaluation Techniques

Boubakeur BELAROUSSI, BioClinica,
boubakeur.belaroussi@bioclinica.com
www.bioclinica.com

Topics

- Objectives
- BioClinica
- Validation in Image Processing
- Validation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

Topics

- Objectives
- BioClinica
- Validation in Image Processing
- Validation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

Objectives

- Set a rigorous evaluation methodology to validate an image processing algorithm
  - Validation framework
  - Choose / Define appropriate datasets
  - Choose / Define the quantitative criteria

Topics

- Objectives
- BioClinica
- Validation in Image Processing
- Validation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

BioClinica

- BioClinica is a **technology-oriented Imaging Contract Research Organization (CRO)** providing biotechnology and pharmaceutical companies with a unique expertise in the field of medical image analysis in the context of multicenter clinical trials.

- BioClinica manages the imaging component of clinical trials (Phase I to IV) using its proprietary image processing software technology. This technology enables the introduction of **quantitative imaging markers** in the design of innovative clinical trials in major diagnostic and therapeutic areas:
  - Central Nervous System (CNS) diseases
  - Neurovascular diseases
  - Vascular diseases
  - Oncology

BioClinica

- The use of accurate and reproducible imaging parameters as safety and efficacy endpoints can dramatically improve the overall quality of clinical trials and lead to more efficient and cost effective drug development strategies.
- BioClinica' services have been designed and optimized to address every technical, organizational and methodological aspect related to the management of medical imaging data in the context of clinical trials:
  - **Design and optimization of imaging protocols (image acquisition and evaluation procedures)**
  - Investigational sites identification, set-up and standardization
  - **Image data centralization and quality control in multicenter contexts**
  - Image processing and parameter extraction using automated procedures
  - **Preparation and conduct of centralized and blinded image review sessions**
  - Web-based real-time trial monitoring
  - **Electronic data management and submission**
  - Medical writing

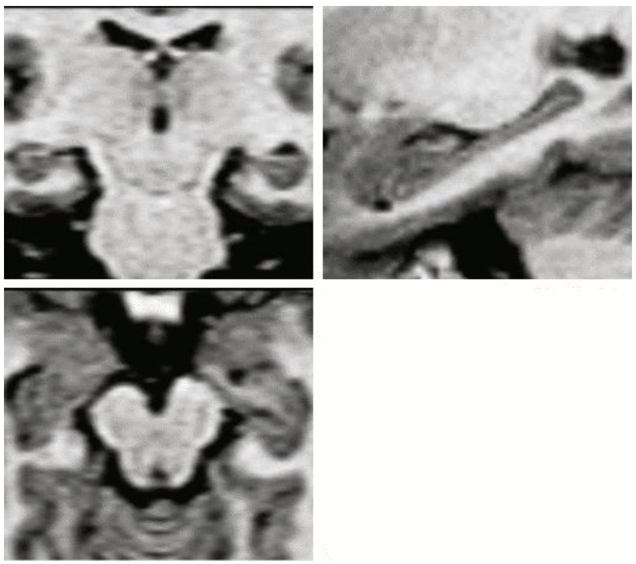
BioClinica

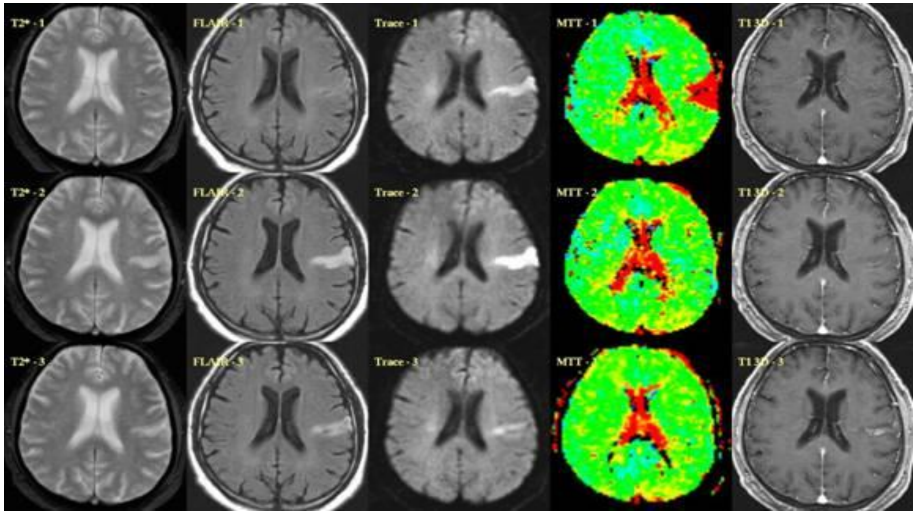
- 3 Sites:
  - Newtown, USA (Headquarters)
  - Leiden, Netherlands (1997)
  - Lyon, France (2007)

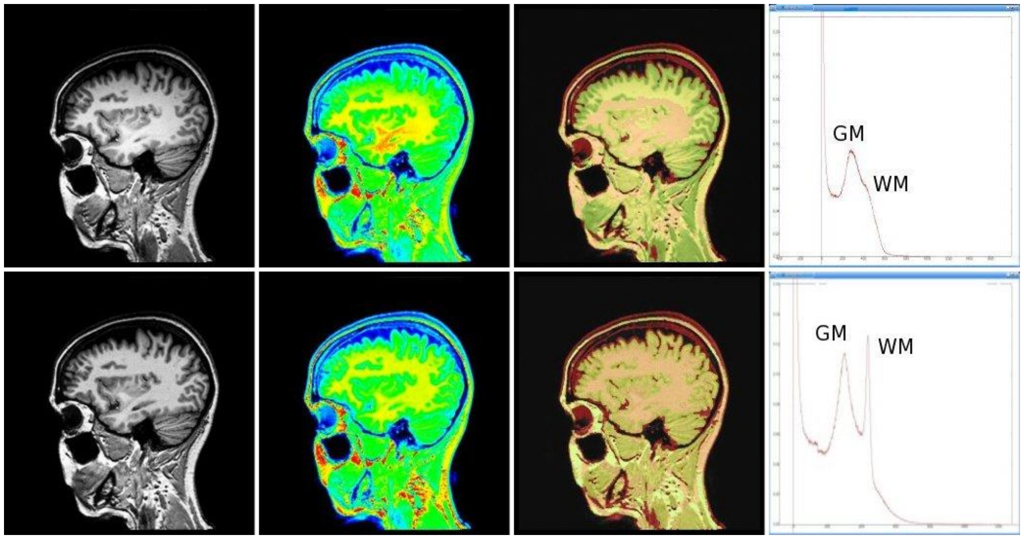
- Imaging Modalities
  - Magnetic Resonance Imaging (MRI)
  - X-Ray
  - DXA
  - TEP
- R&D Projects (Image Processing)
  - 3D Statistical Segmentation
  - 3D Registration
  - 3D Hippocampus Detection
  - 3D Brain/Hippocampal Atrophy Measurement
  - Atlas-based segmentation

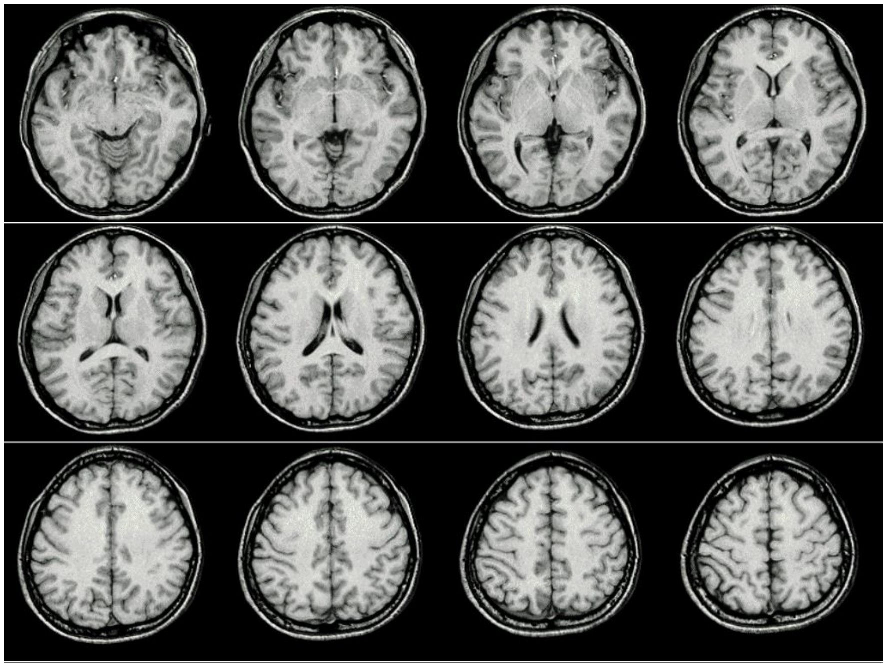
Bio-Imaging Technologies

- Illustrations









Bio-Imaging Technologies

- Our customers
  - Pharmaceutical Groups
    - Novartis
    - Servier
    - Roche
    - Wyeth
  - Academic
    - HCL
    - CHU Bordeaux
    - Pitié Salpétrière

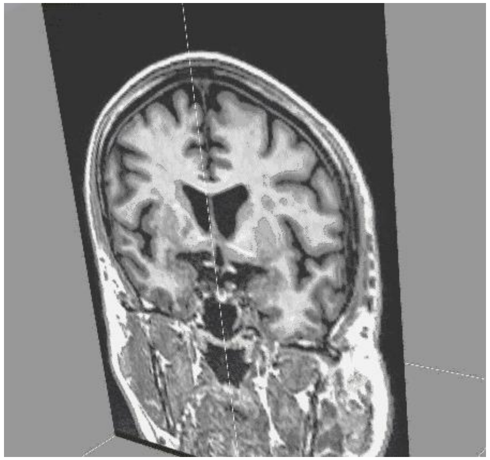
- Example of clinical trials
  - Cushing disease
  - Breast Oncology
  - Alzheimer's Disease
  - Multiple Sclerosis

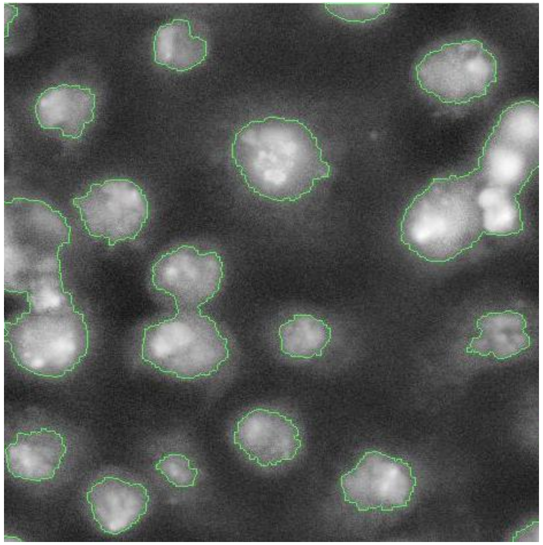
Topics

- Objectives
- BioClinica
- Validation in Image Processing
- Validation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

Validation in Image Processing

- **Increasingly important role of image processing in many fields (medicine, computer vision, telecommunication ...)**
  - The performance of image processing methods may have an important impact on the performance of the larger systems as well as on the human observer that needs to analyze all of the available image data
  - Sources of error are numerous in image processing





Validation in Image Processing

Importance of Validation in Image Processing

- Validation of image processing methods is required to:
  - Understand and highlight the intrinsic characteristics and behaviour of a method,
  - Evaluate performance and limitations,
  - Eventually to compare these performances with different methods.

- Results of validation studies help in improving image-processing performances

- Algorithmic advances in image processing are often stimulated by the recognition of the need for an image analysis capability that does not yet exist.

Validation in Image Processing

Importance of Validation in Image Processing

- The characteristics of the need, such as the ultimate requirements for accuracy or for speed, and the type of images under consideration, provide constraints on the algorithm and its implementation.

- Validation strategies then provide the essential assessment by which any particular algorithm and its implementation will be judged as acceptable or unacceptable, given the constraints of the particular image analysis challenge to be addressed.

- **Although algorithm development alone is often the contribution of research in this area, it is not possible to create algorithms that will have a significant impact in clinical practice, without simultaneously considering the validation in the context of the problem constraints.**

Validation in Image Processing

Challenges in Validation

- Further research is needed in validation for image processing as issues concerning validation are numerous.
- Mathematical and statistical tools are required for quantitative evaluation or for estimating performances in the absence of a suitable « ground truth », « gold standard » or other reference standard
- Comparison of the performance of different methods requires the use of standardized or at least a rigorous terminology and common methodology for the validation process
- I am convinced that general frameworks or validation guidelines could be established to improve validation in image processing.
- Validation data sets with available Ground Truth are required.

Validation in Image Processing

Challenges in Validation

- Validation is rarely the main objective of traditional papers in image processing.
- Innovation usually stands in the image processing methods itself and validation is usually addressed only as a section in the paper.
- However, validation is by itself a research topic where methodological innovation and research are required.

Validation in Image Processing

Conclusion

- The development of standards for terminology methodology and data sets used in evaluation
- The ability to create test data sets and evaluation metrics that capture the critical features of important classes of image analysis problems, and so enable generalizable conclusions to be drawn about the efficacy of particular analysis methods.
- The study of cumulative performance and error propagation along complex image processing workflows. Quite often, a processing technique is developed and validated for essentially a single point in the time, but images are often used in other ways once processed. For example, the performance of computer-aided detection and diagnosis (CAD) schemes generally depends critically on the state of the image data being input to them.
- Extension of validation techniques to other lesion categories and other types of images and / or modalities. Many image processing techniquesand thus the approaches used to validate them are often designed for specific lesion types in specific types of images. Ways to generalize these techniques need to be explored.

Context

- Objectives
- BioClinica
- Validation in Image Processing
- Evaluation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

Evaluation Methodology

6 months for R&D projects

3 years to obtain a Ph.D degree

Objective

↓

Design / implementation: (UML, object-oriented ...)

↓

Software / Implementation Validation: (Unit Tests, black box tests...)

↓

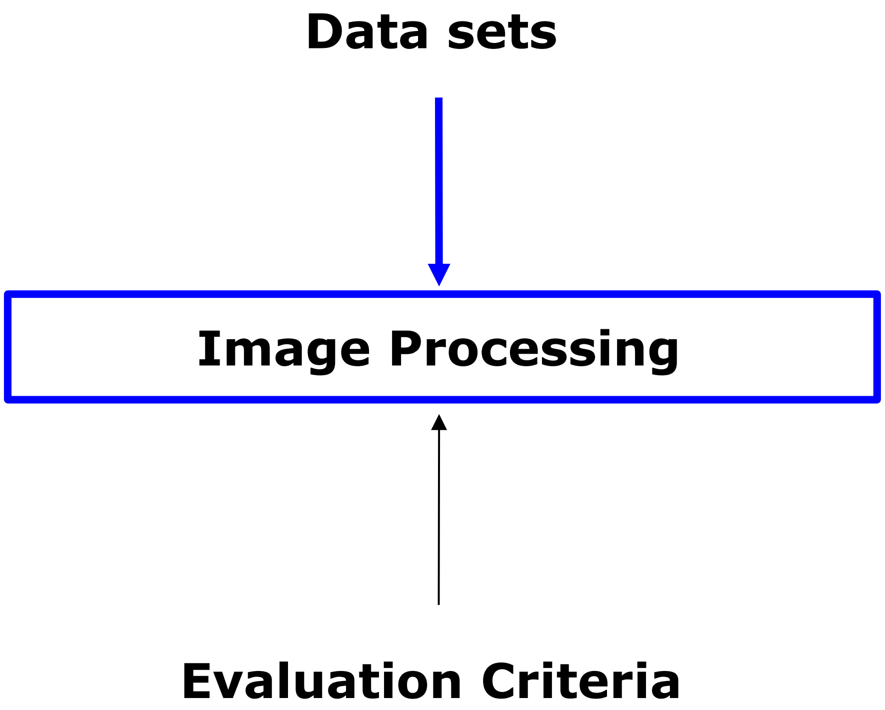
Image Processing Validation

↓

Application Validation: (GUI, batch mode, reproducibility....)

Evaluation Methodology

Image Processing Validation Workflow

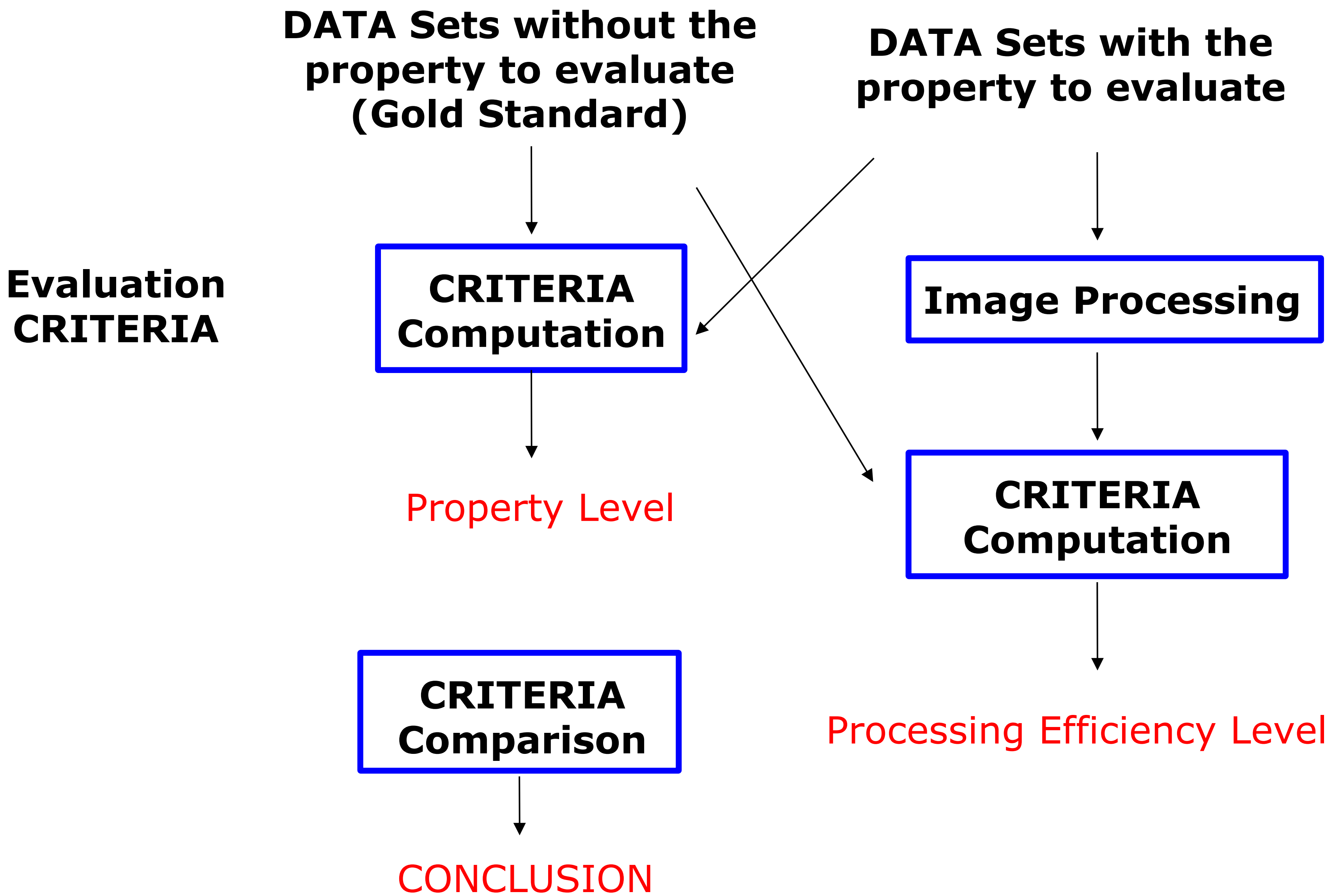


Evaluation Methodology

- Once an image processing technique has been implemented

1. Choose / Define relevant **DATA** with/without the property to evaluate (context-dependent),
2. Choose / Define relevant **Evaluation Criterion (ia)** that allow you the quantification of the property to evaluate,
3. Compute evaluation criterion (ia) on the **DATA** prior to applying the proposed image processing technique,
4. Process the **DATA** using the proposed image processing technique,
5. Compute evaluation criterion (ia) on the processed **DATA**
6. Compare criterion (ia) value(s) prior to / after processing
7. Intermediate/Final Conclusion

Evaluation Methodology



Evaluation Methodology

A 3-Levels Evaluation <u>DATA</u>

Simple Synthetic DATA GOLD STANDARD ?

Intermediate conclusion

Realistic Synthetic DATA

Intermediate conclusion

Real DATA GOLD STANDARD ?

Conclusion

Context

- Objectives
- BioClinica
- Validation in Image Processing
- Evaluation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

Image Filtering

Definition

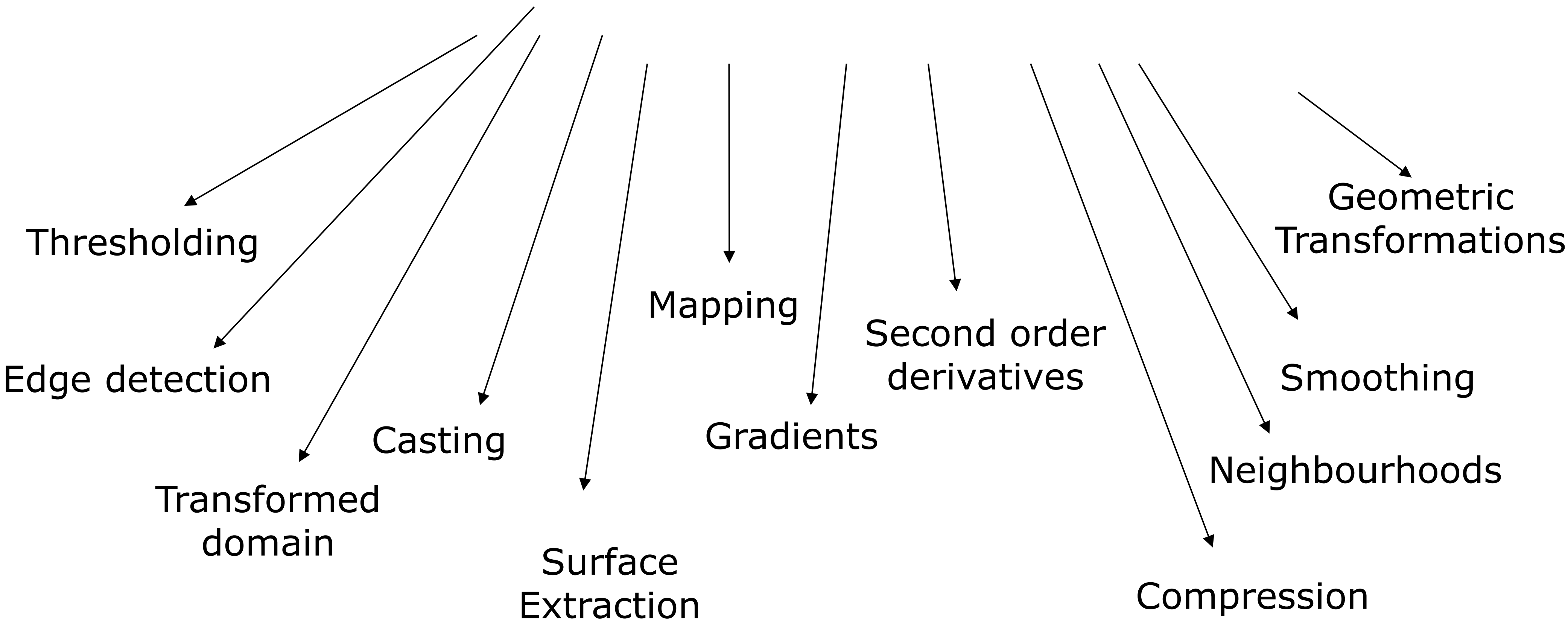
- Processing on an image performed by combining or comparing individual pixels with their neighbours.

Applications

- Most of the time, used as pre-processing step
- Many interesting and useful effects can be obtained, such as:
  - Sharpening,
  - Blurring,
  - Edge detection,
  - Embossing,
  - Compression ...

Image Filtering

▶ Overview of Filtering techniques



How to Evaluate Filtering efficiency ?

Image Filtering Evaluation Criteria

Mean Square Error (MSE)

- Measures the average of the square error

$$MSE(u,v) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left| \, u(m,n) - v(m,n) \, \right|^2$$

- A lower value of MSE signifies lesser error
- Heavily weight outliers !!

Root Mean Square Error (RMSE)

- Quantifies the average sum of distortion in each pixel of the reconstructed image

$$RMSE = \sqrt{MSE}$$

Image Filtering Evaluation Criteria

- Mean Absolute Deviation (MAD) Distance
  - Is the average of the absolute deviations and is a statistical summary of variability
  - Is a common measure of forecast error in time series analysis

$$MAD(X, Y) = E(|x - y - E(X, Y)|)$$

Image Filtering Evaluation Criteria

Signal to Noise Ratio (SNR)

- Defined as the ratio of the mean pixel value to the SD of the pixel values
- Measures the performance of lossy compression algorithms
- Can be computed on a given Region of Interest or on the whole image
- For a stochastic signal (S and N are independent):

$$SNR = 20.\log_{10}\left(\frac{\mu_{Signal}}{\sigma_{Noise}}\right)$$

Image Filtering Evaluation Criteria

Peak Signal to Noise Ratio (PSNR)

- Measures the estimates of the quality of reconstructed image compared with an original
- Is a standard way to measure image fidelity
- Is a single number (in dB) that reflects the quality of reconstructed image

$$PSNR = 20\log_{10}\left(\frac{S}{RMSE}\right)$$

- Where S is the maximum pixel value (S=255 for UCHAR images)
- Higher value of ¨ PSNR is better !

Image Filtering Evaluation Criteria

Correlation Coefficient (Pearson's correlation)

- Quantifies the closeness between two images

$$Corr(A/B) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(A_{i,j}-\overline{A})(B_{i,j}-\overline{B})}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(A_{i,j}-\overline{A})^2\sum_{i=1}^{M}\sum_{j=1}^{N}(B_{i,j}-\overline{B})^2}}$$

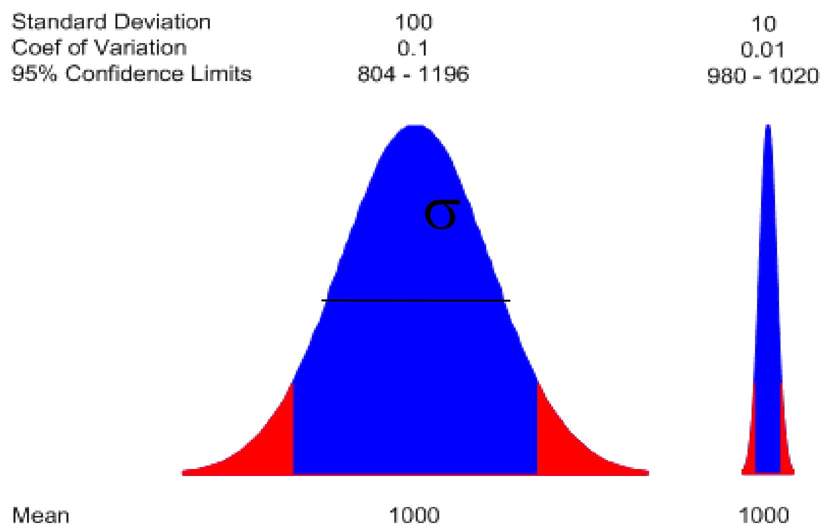
- Values range from -1 to +1
- 1 indicates that the images are exactly the same
- -1 indicates that the images are exactly opposite to each other
- 0 indicates that the images are not correlated

Image Filtering Evaluation Criteria

Coefficient of variation

- Is a measure of dispersion of a probability distribution
- Dimensionless number (or % if x100)
- Defined inside a region of interest (White Matter mask ...)
- Only characterizes within-class scattering

$$c_v = \frac{\sigma}{\mu}$$



- When the mean value is close to zero, CV is sensitive to change in the SD

Image Filtering Evaluation Criteria

Coefficient of joint variations

- Characterizes between-class scattering

$$CJV(class1, class2) = \frac{(\sigma(class1) - \sigma(class2))}{(|\mu(class1) - \mu(class2)|)}$$

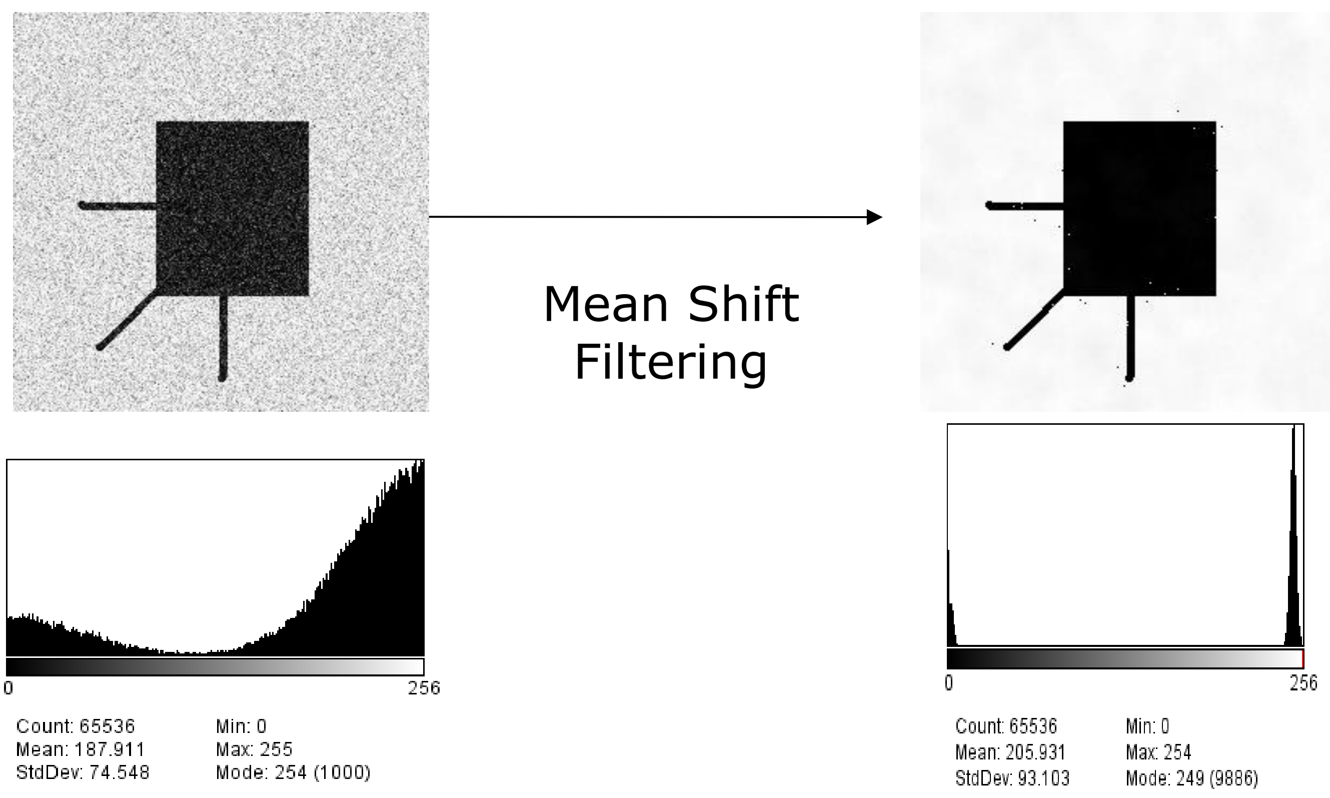


Image Filtering Evaluation Criteria

Quality Index (Wang_02)

- Models Image Distortion as a combination of 3 factors

$$Q = \underbrace{\frac{\sigma_{xy}}{\sigma_x \sigma_y}}_{\substack{\text{Correlation} \\ \text{Coefficient} \\ [-1;1]}} \cdot \underbrace{\frac{2\,\bar{x}\,\bar{y}}{(\bar{x})^2 + (\bar{y})^2}}_{\substack{\text{Mean luminance} \\ \text{closeness} \\ [0;1]}} \cdot \underbrace{\frac{2\,\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}}_{\substack{\text{Contrast} \\ \text{Similarity} \\ [0;1]}}$$

[-1;1]

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Image Filtering Evaluation Criteria

Image Entropy (Shannon)

- Measures the amount of disorder in a system

n grey levels

$$H(X) = -\mathbf{E}[\log_2 p(i)] = \sum_{i=1}^{n} p(i) \log_2 \left(\frac{1}{p(i)} \right) = -\sum_{i=1}^{n} p(i) \log_2 p(i).$$

Image Filtering Evaluation Criteria

- Example of Criteria Comparison (MSE and Q)

(a) (b) (c) (d)

Image Quality ?

Image Filtering Evaluation Criteria

Example of Criteria Comparison (MSE and Q)

a) Original Image
b) Salt-pepper noise
c) Gaussian noise
d) Speckle noise

b) MSE=225 - Q=0.6494

c) MSE=225 - Q=0.3891

b) MSE=225 - Q=0.4408

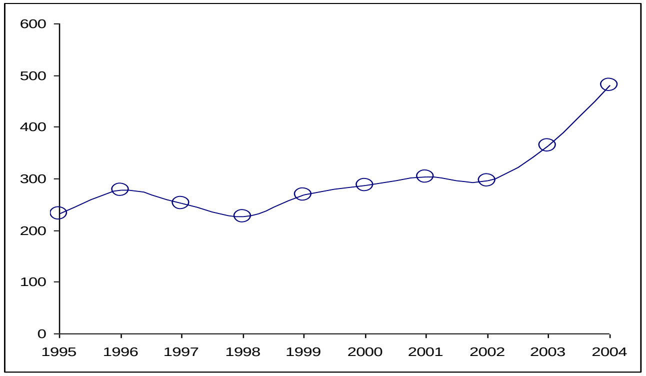
Image Segmentation

- Definition
  - Process of partitioning an image into multiple regions (sets of pixels)
- Applications
  - Medical Imaging
  - Object Recognition
  - Computer graphics
  - Airport security systems
  - Robotic vision

Image Segmentation

- The choice of a segmentation technique depends on
  - The image (light, noise, texture ...)
  - The final objective (qualitative, measure, interpretation ...)
  - Primitives to extract (edges, regions, textures ...)
  - The operating constraints (real-time, complexity ...)

Number of publications →



Source : Elsevier

- Existing surveys
  - Surveys regarding given methods (level set, ...)
  - Specific applications (MRI segmentation ...)
  - No surveys regarding all the techniques !

Image Segmentation

- Overview of Segmentation techniques

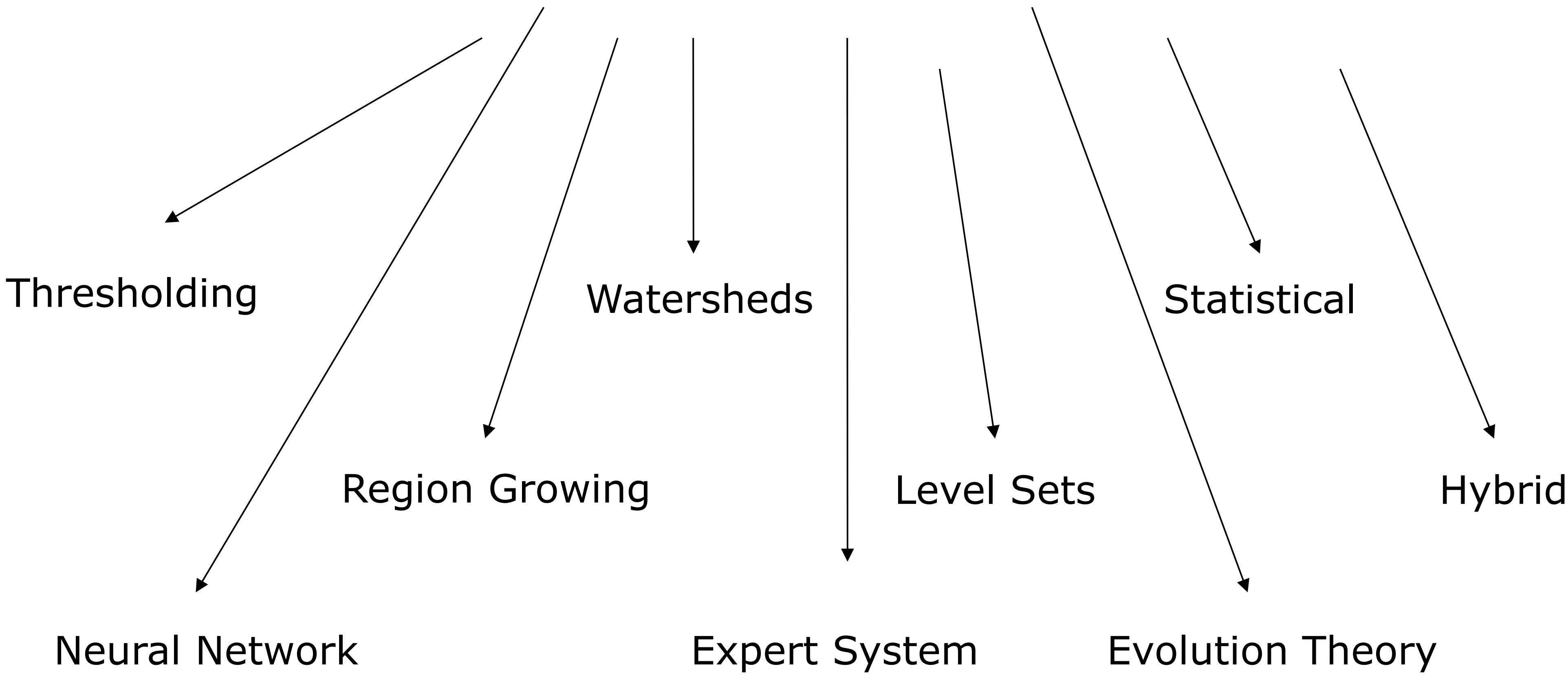


Image Segmentation

- Findings
  - No general theory for image segmentation
  - Many segmentation techniques
- How to choose the appropriate algorithm ?
  - Need to evaluate the segmentation techniques
- Three evaluation levels
  - Optimizing a method
  - Comparing methods
  - Evaluate acquisition impact

Image Segmentation

Segmentation Characterization

- The purpose of evaluation for a specific algorithm is to quantitatively recognize its behavior in treating various images and/or to help appropriately setting its parameters regarding different applications to achieve the best performance of this algorithm
- This process could also help to improve the functioning of the algorithm under consideration
  - Giving different values to the algorithm's parameters for segmenting some comparable images and then evaluating the influence of multiple settings of the algorithm over its performance. The adaptability and the best performance of this algorithm for given images are evaluated.
  - Using the same parameter setting of the algorithm for segmenting multiple images. The ability and consistency of the algorithm in treating images with different contents and/or acquired under various conditions are evaluated.

Image Segmentation

- Overview of Evaluation techniques

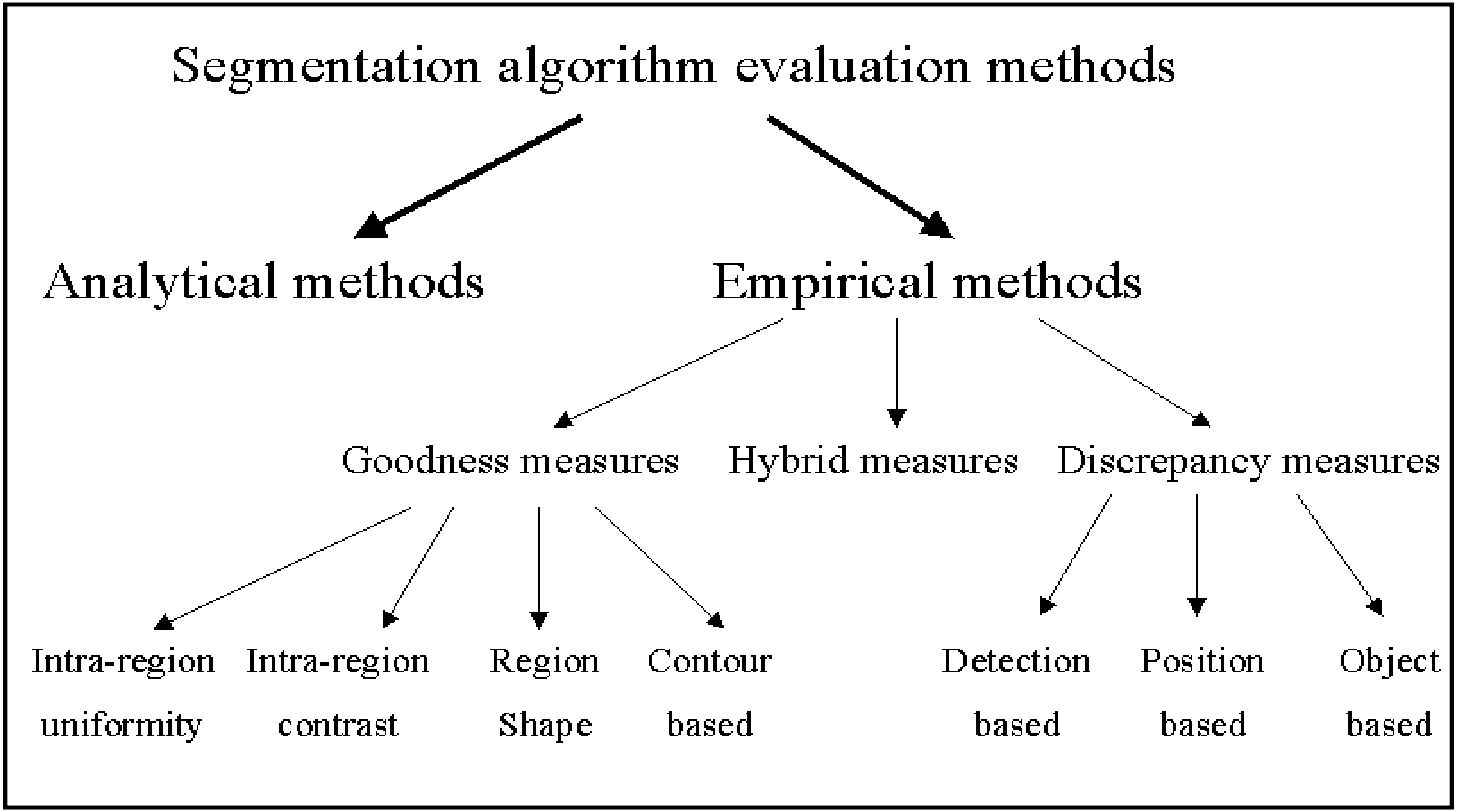


Image Segmentation

Analytical Methods

- **Consider the algorithm itself**
  - Underlying theory (levels set, thresholding ...)
  - Amount of a priori knowedge incorporated into the algorithm
  - Processing strategy: Parallel, Sequential ...
  - Processing complexity and efficiency
  - Probability of correct detection / probability of false detection
  - Resolution of segmented images: pixel, sub-pixel, group of pixels...

- Very hard to apply today due to the large

Image Segmentation

- Analytical Methods Illustration (Canny Edge detection)

Optimal edge detection algorithm

- Good Detection
  - The algorithm should mark as many real edges in the image as possible
- Good Localization
  - Edges marked should be as close as possible to the edge in the real image
- Minimal Response
  - A given edge in the image should only be marked once, and where possible, image noise should not create false edges

Image Segmentation

Empirical Methods

- **Consider the result of the algorithm**

- *Empirical Goodness methods*
  - Unsupervised,
  - Standalone: do not require a reference image,
  - Subjective: evaluate a segmentation based on how well they match a broad set of characteristics as desired by humans

- *Empirical Discrepancy methods*
  - Supervised, relative, objective

- *Empirical Hybrid methods*
  - Combine characteristics of both Discrepancy and Goodness methods

Empirical Goodness methods

- What is a **“good”** segmentation? (Haralick)
  - Regions should be uniform and homogeneous with respect to some characteristics
  - Adjacent regions should have significant differences with respect to the characteristic on which they are uniform
    - ***Characteristics criteria***
  - Region interiors should be simple and without holes
  - Boundaries should be simple , not ragged, and be spatially accurate
    - ***Semantic criteria***

Empirical Goodness Methods

- Uniformity Measure (Lévine and Nazif 1984)

Segmentation : $\{R_i, i=1,2\ldots R\}$ $A = \sum_i A_i$ A_i = area (R_i)

$\sum_i \frac{\sigma_i^2}{\sigma_{\max}^2}$ → *Region homogeneity*

Empirical Goodness Methods

- Liu and Yang (1994)

Segmentation : $\{R_i, i=1,2\ldots R\}$ $$A = \sum_i A_i$$ A_i = area (R_i)

$$Q = \frac{1}{1000(N \times M)} \sqrt{R} \sum_i \frac{e_i^2}{\sqrt{A_i}}$$

→ *Region homogeneity*

→ *Region simplicity (no holes)*

→ *Contrast between regions*

Biases:

1. Segmentations with lots of regions are heavily penalized (R)

2. Segmentations that have regions with large areas are heavily penalized

Empirical Goodness Methods

- Borsotti et al. (1998)

→ *Liu et al. Measure improving*

Segmentation : $\{R_i,\ i=1,2\ldots R\}$ $\qquad A=\sum_i A_i \qquad$ A_i = area (R_i)

$$Q=\frac{1}{10000\,A}\sqrt{R}\sum_{i}\left[\underbrace{\frac{e_i^2}{1+\log A_i}}_{1}+\underbrace{\left(\frac{R(A_i)}{A_i}\right)^2}_{2}\right]$$

1 → *Penalises non homogeneous regions (large regions)*

2 → *Penalises the large presence of a number of regions with the same area (holes)*

Empirical Goodness Methods

- Entropy-based Goodness method (Zhang et al. 2004)
  - A good segmentation evaluation should
    - Maximize the uniformity of pixels within each segmented regions
    - Minimize the uniformity across the regions
  - Consequently, **entropy**, a measure of the disorder within a region is a natural characteristic to incorporate into a segmentation evaluation method

Empirical Goodness Methods

- Entropy-based Goodness method (Zhang et al. 2004)

Region Entropy

- Expected entropy across all regions where each regions has weight (or probability) proportional to its area
- Used as a measure of uniformity within the regions of I

$$H_r(I) = \sum_{j=1}^{N} \left(\frac{S_j}{S_I}\right) H(R_j)$$

S_j = area (R_j)

$H(R_j)$= Entropy (R_j)

Layout Entropy

- Encodes a representation for the segmentation

$$H_\ell(I) = -\sum_{j=1}^{N} \frac{S_j}{S_I} \log \frac{S_j}{S_I}$$

Effectiveness Measure

$$E = H_\ell(I) + H_r(I)$$

Empirical Goodness Methods

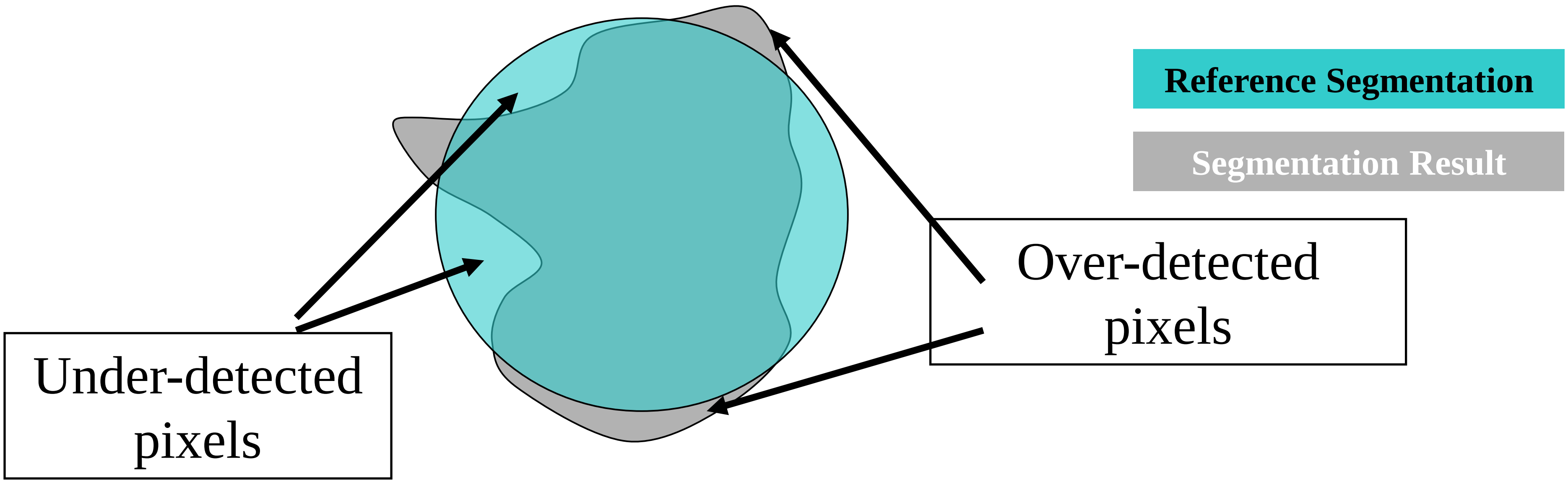
- Empirical Goodness Methods Conclusion / Future directions
  - Perform reasonably-well in evaluating different segmentation results produced by the same algorithm
  - But, more modest performance in comparing segmentation results produced by different algorithms and in comparing human versus machine segmentation
  - 4 major problems to address in the future
    - The existing intra-region uniformity metrics are too sensitive to noise and are biased towards under-segmentation,
    - Most existing metrics assume a single underlying distribution, usually Gaussian-like, of pixels in a segment,
    - The homogeneity and disparity metrics are frequently not balanced and do not complement each other effectively,
    - All the evaluation methods use only low-level features and do not incorporate semantic information

Empirical Discrepancy methods

- What is a **"discrepancy"** ? (Canny)
  - Good Detection
    - The algorithm should mark as many real edges in the image as possible
    - ***Detection criteria***
  - Good Localization
    - Edges marked should be as close as possible to the edge in the real image
    - ***Localization criteria***

Empirical Discrepancy methods

- Measure the amount of agreement between a segmentation result to a reference segmentation
  - Correctly detected pixels: True positive (TP) and True negative (TN)
  - Over-detected pixels: False positive (FP)
  - Under-detected pixels: False negative (FN)



Detection Criteria

Detection Rates

- **Sensitivity**

$$p = \frac{TP}{(TP + FN)}$$

- **Specificity**

$$q = \frac{TN}{(TN + FP)}$$

- **Prevalence**

$$\pi = \frac{(TP + FN)}{(TP + FP + TN + FN)}$$

- **Level of test**

$$\theta = \frac{(TP + FP)}{(TP + FP + TN + FN)}$$

$$0 \leqslant Rate \leqslant 1$$

Detection Criteria

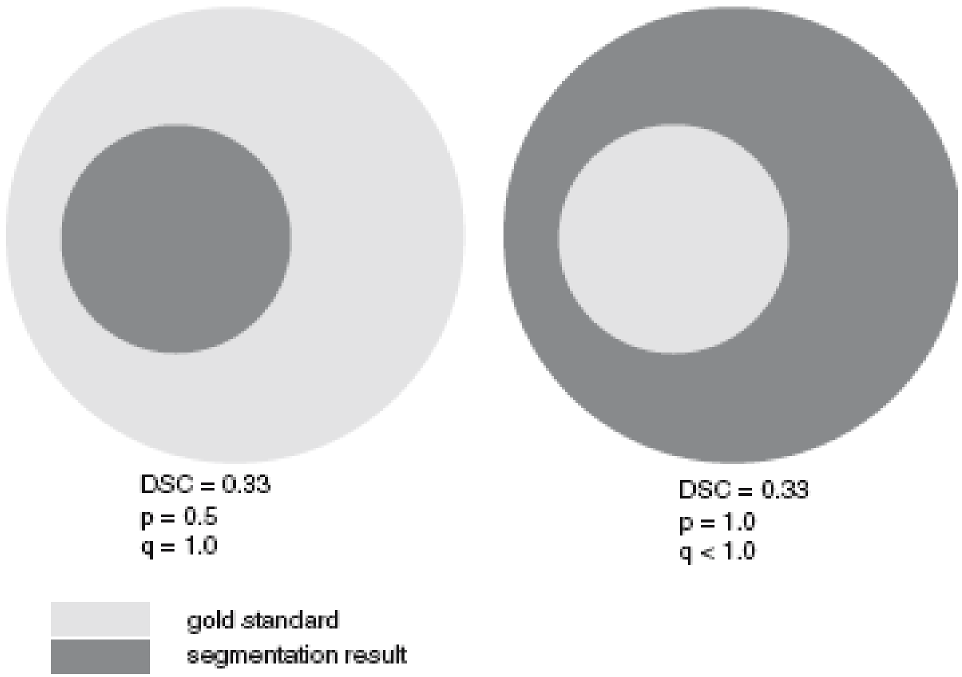
Detection Rates

$$0 \leqslant DSC \leqslant 1$$

Dice Similarity Measure (DSC) $$DSC = \frac{2.TP}{(2.TP + FP + FN)}$$

- Amount of the intersection between a segmented object and the gold standard (Positive Specific Agreement)

⑤ Although geometrically intuitive, it lacks the information about the type of segmentation error

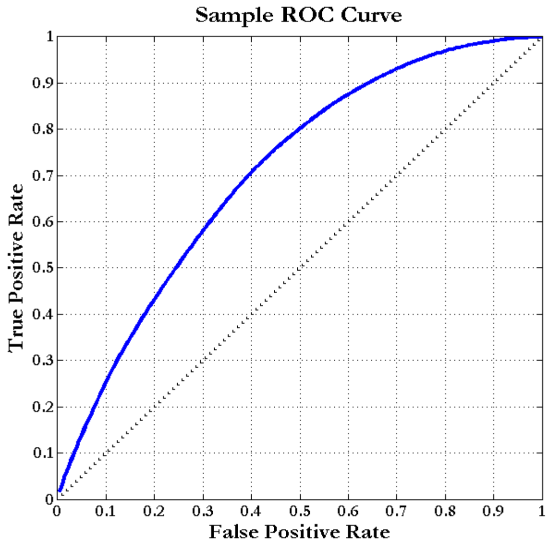


Detection Criteria

Detection Rates

Receiver Operating Characteristics (ROC) **Curve**

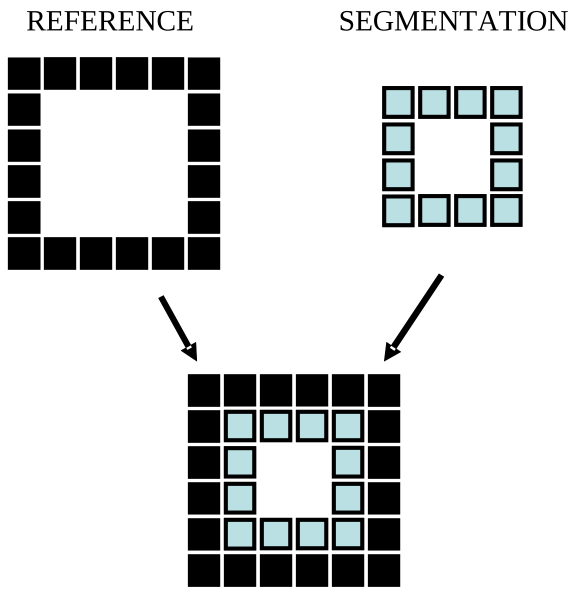
- Is a graphical plot of the **sensitivity** versus (**1- specificity**) for a binary classifier as its discrimination threshold is varied
- In ROC analysis, the Area Under Curve (AUC) defines to which amount the classifier under investigation is better than a random classifier (AUC being 0.5)



Detection Criteria

Detection Rates

- ***Counter-example***



Detection rates values

- p = 0 %

- DSC = 0%

Does it mean that the detected contour is incorrect?

Detection Rates are not sufficient, localization criteria MUST be taken into account !

Localization Criteria

Yasnoff Distance (Yasnoff_79)

- ***First distance-based criterion proposed in the context of image segmentation evaluation***

- ***Distance between a FP pixels and its nearest (euclidean distance) pixels in the reference segmentation***

$$\frac{100}{A} \times \sqrt{\sum_{s} d^2(s)}$$

Euclidean distance

Area

Localization Criteria

- Figure Of Merit (Pratt_79)

$$FOM = \frac{1}{M}\sum_{i=1}^{N_e - N_h + N_b} \frac{1}{1 + \alpha\, d(i)^2}$$

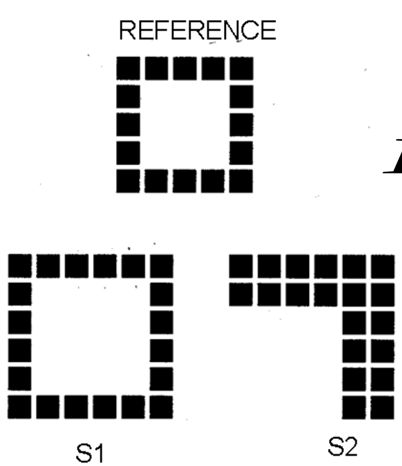
$$M = \max(N_e, N_e - N_h + N_b)$$

Ne: Number of pixels in the reference contour
Nb: Number of over-detected pixels
Nh: number of under-detected pixels

d(i): euclidean distance between a declared edge of the detected contour and the nearest reference edge pixel

$$FOM \in [0;1]$$

<u>Counter-example</u>



$$FOM(S1) = FOM(S2)$$

Does not consider Error shape

Does not consider under-segmentation

Localization Criteria

▶ Hausdorff Distance (Huttenlocher_93)

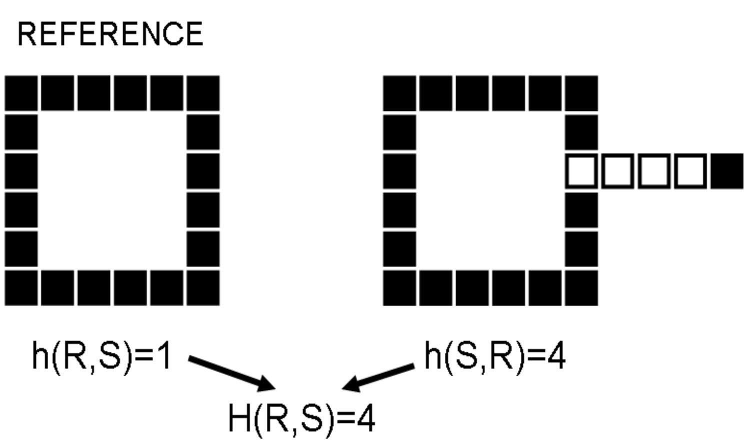
⑤ ***Maximum distance of a set to the nearest point in the other set***

$$H(R,S) = \max(h(R,S), h(S,R))$$

$$h(R,S) = \max_{r \in R} \min_{s \in S} \| r - s \|$$

Euclidean distance

<u>Counter-example</u>



Very sensitive to noise !

Localization Criteria

- Baddeley Distance (Wilson_97)
  - Consider both pixel **position** and pixel **intensity**
  - A: reference and B: segmentation result
  - p: hyperparameter to differently weight the error (p>=1)

$$D_H^p(A,B) = \sqrt[p]{\frac{1}{Card(A \cup B)} \sum_{x \in (A \cup B)} (d(x,B) - d(x,A))^p}$$

Euclidean distance

Detection Criteria

Scalable Discrepancy Measures (Belaroussi_02)

- Consider both under (UD) and over detected (OD) pixels within an adjustable area (d_{TH})
- Give discrepancy intensity (I) and its relative position (P)
- A scale parameter (n) allows the measures accuracy adjustment

Euclidean distance

Over-Detection Discrepancy

- Intensity

$$ODIn = \frac{1}{N_O} \cdot \sum_{i=1}^{No} \min\left\{ \left(\frac{d_O(i)}{d_{TH}} \right)^n, 1 \right\},$$

- Relative Position

$$ODRPn = \frac{1}{N_O} \cdot \sum_{i=1}^{No} \min\left\{ \left(\frac{d_O(i)}{d_{TH}} \right)^n, 1 \right\} * sign(d_O(i)),$$

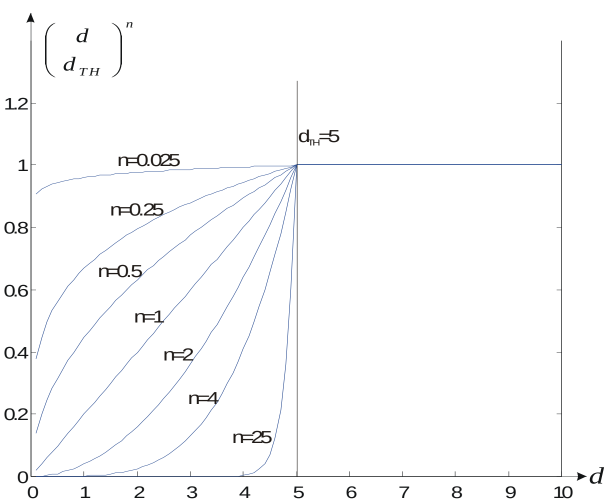
Under-Detection Discrepancy

- Intensity

$$UDIn = \frac{1}{N_U} \cdot \sum_{j=1}^{Nu} \min\left\{ \left(\frac{d_U(j)}{d_{TH}} \right)^n, 1 \right\},$$

- Relative Position

$$UDRPn = \frac{1}{N_U} \cdot \sum_{j=1}^{Nu} \min\left\{ \left(\frac{d_U'(j)}{d_{TH}} \right)^n, 1 \right\} * sign(d(j)),$$



Detection Criteria

Scalable Discrepancy Measures (Belaroussi_02)

- Comparison to similar metrics

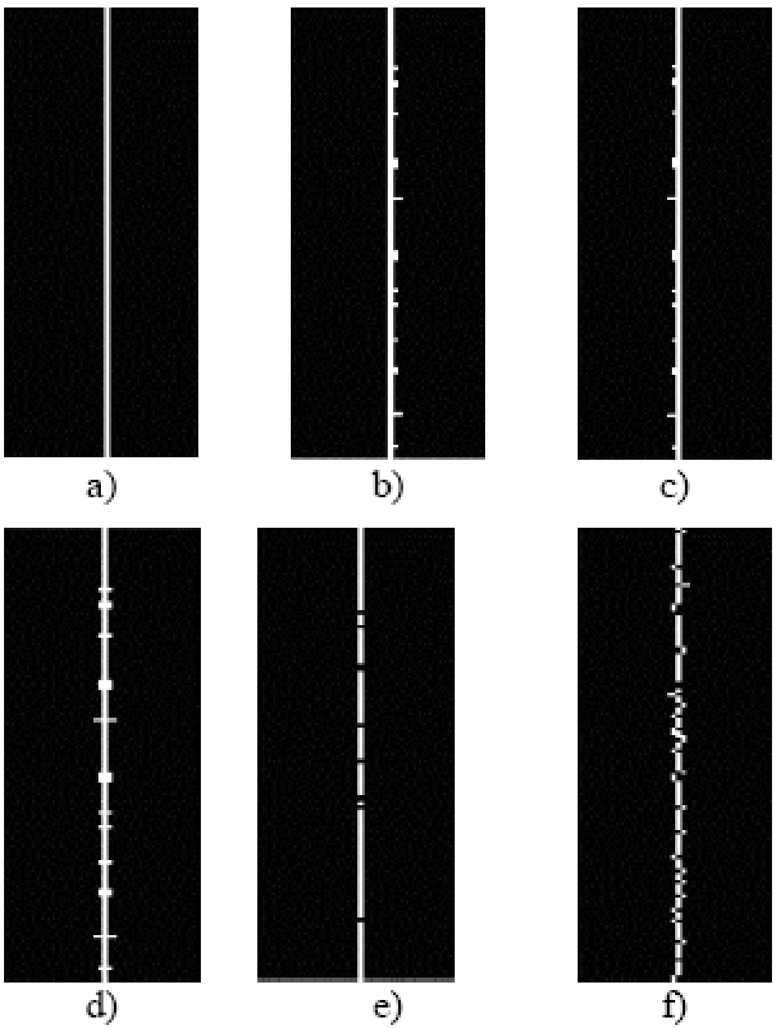
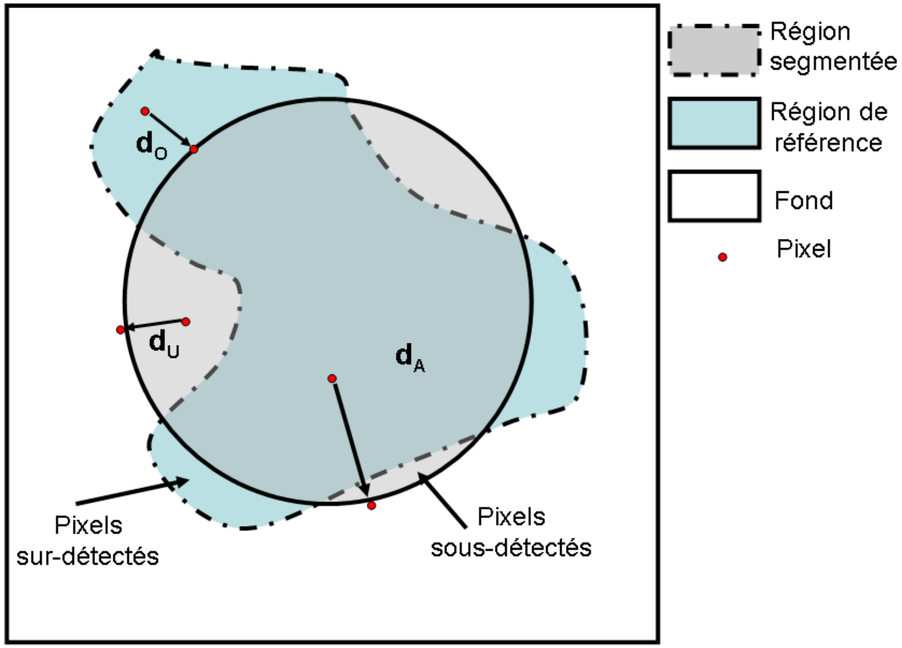


Image	a)	b)	c)	d)	e)	f)
ODI_n	0	0.052	0.052	0.052	0	0.048
ODP_n	0	-0.052	0.052	0	0	0
UDI_n	0	0	0	0	0.04	0.04
UDP_n	0	0	0	0	-0.04	-0.032

Discrepancy criterions	a)	b)	c)	d)	e)	f)
Pratt's FOM (α=1/9)	0	0.14	0.14	0.24	0	0.23
Hausdorff Distance	0	4	4	4	1	4

Detection Criteria

- New Discrepancy Measures (Goumeidane_03)
  - Consider the compactness of the region under investigation
  - Internal Distortion Rate (IDR) for under-detection
  - External Distortion Rate (EDR) for over-detection



Euclidean distance

$$IDR = \frac{\sqrt{\sum_{i=1}^{K1} d_u^2(i)}}{\sqrt{\sum_{j=1}^{N} d_A^2(j)}}$$

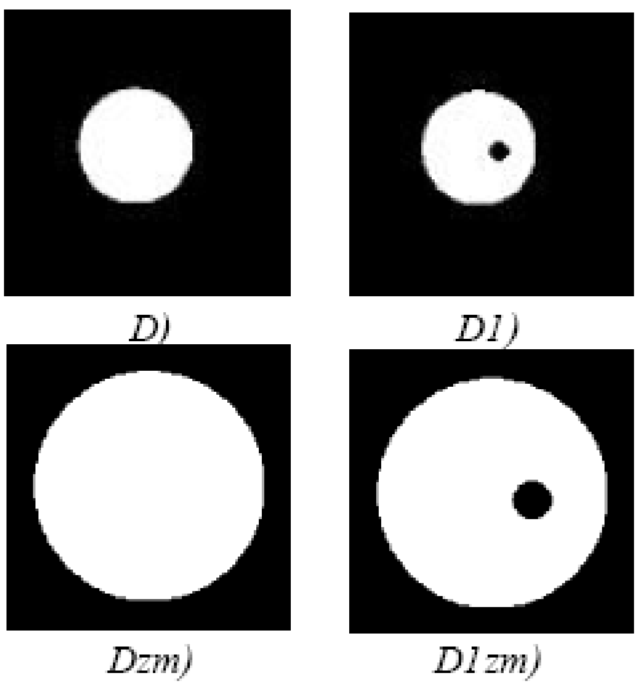
$$EDR = \frac{\sqrt{\sum_{i=1}^{K2} d_o^2(i)}}{\sqrt{\sum_{j=1}^{N} d_A^2(j)}}$$

$$PSR = IDR + EDR$$

Detection Criteria

New Discrepancy Measures (Goumeidane_03)

- Comparison to similar metrics



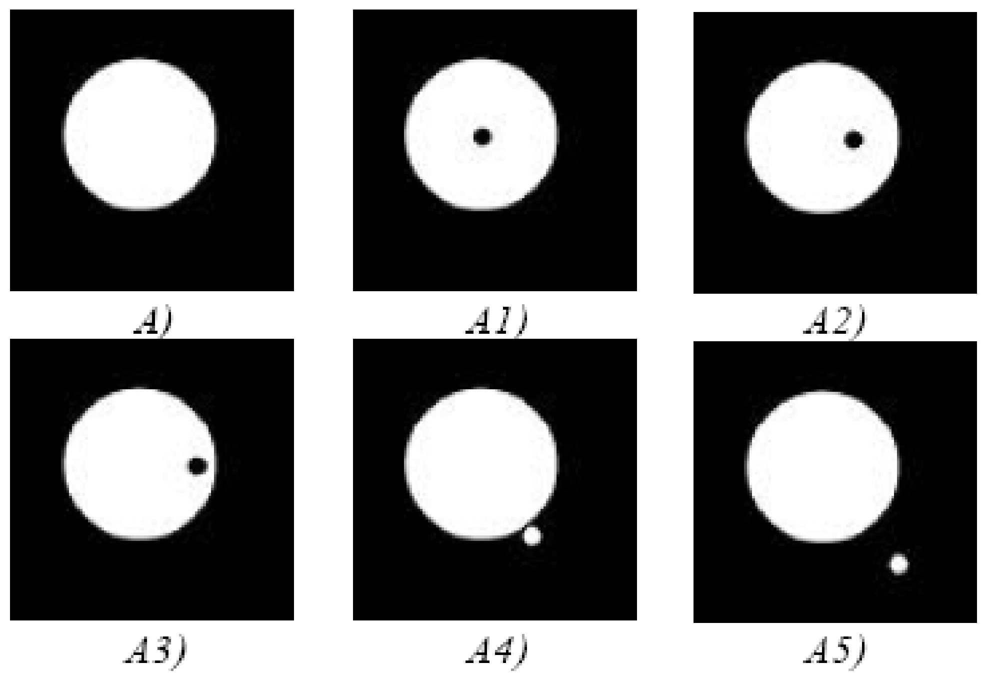


Image	*IDR* (%)	*EDR* (%)	*YasD* (%)
D1	24,89	0	0.79
D1zm	24.92	0	3.14

Image	*IDR* (%)	*EDR* (%)	*YasD* (%)	*BdD*
A1	26.41	0	1.47	0.24
A2	17.17	0	0.95	0.24
A3	7.29	0	0.4	0.24
A4	0	6.81	0.38	8.93
A5	0	25.21	1.4	20.38

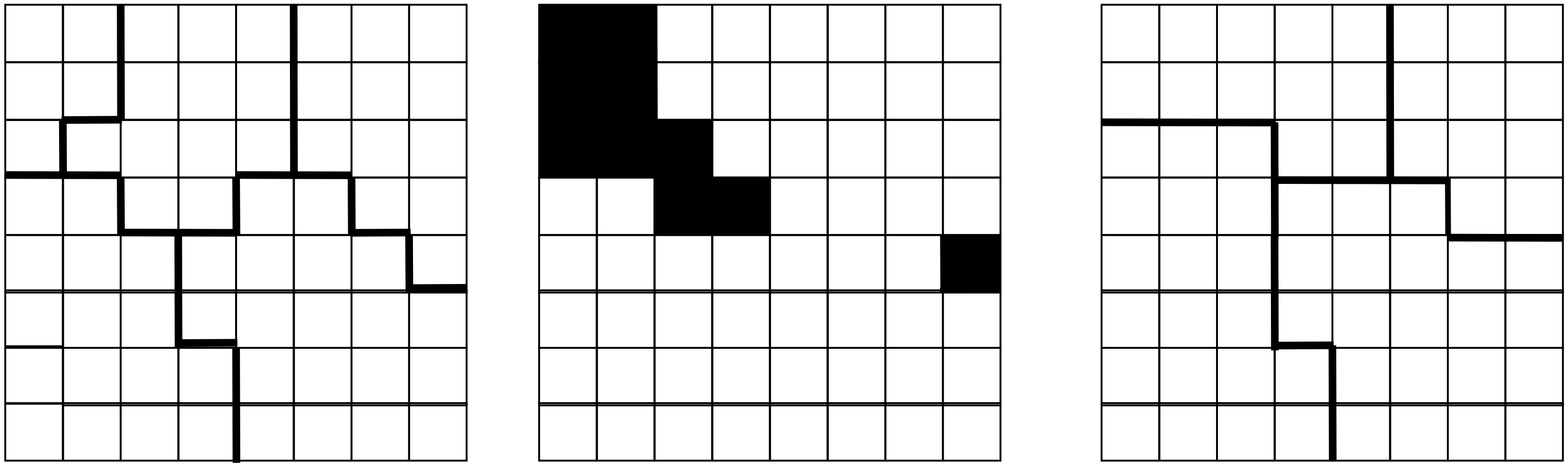
Detection Criteria

- Partition Distance (Cardoso_05)
  - Image Segmentation ->Partitions creation from a given image
  - Segmentation Evaluation -> Partitions comparison
  - **Metric**: ***Partition Distance*** (Almudevar_99)
    - R is a given image of N pixels
    - A partition of R is a set of exclusive clusters
    - Let **P** and **Q** being 2 partitions of R
    - Partition Distance is the minimum number of pixels to shift in clusters of **P** so that the new clusters set fit the clusters in **Q**.
    - Assignment problem solved using the Hungarian Algorithm (combinatorial optimization algorithm)

Detection Criteria

Partition Distance

- Illustration



Partition A   Les pixels à déplacer   Partition B

- In the whole image, 64 (8x8) pixels

- The 2 partitions have 10 misclassified pixels

- $D_{partition} = 10/(64-1) = 0,16$

Hybrid Methods

Hybrid Methods (Roman-Roldan_01)

- Consider both goodness and discrepancy of the segmented result compared to a reference segmentation
- Discrepancy
  - Sum of both over-detection and under-detection errors
- Goodness
  - Consider error shape and are biased towards under-segmentation,
  - Neighbourhood: over a given distance (3x3x3 window), the current error is no more considered
  - Error interaction: shape errors will be differently weighted

$$R = a\sum_{FP} e_{FP} + b\sum_{FN} e_{FN}$$

Hyperparameters (a,b) are obtained from a training dataset and subjective evaluation

Surface Comparison

Computer Graphics Field

- Remeshing techniques Comparison
- Triangle Mesh

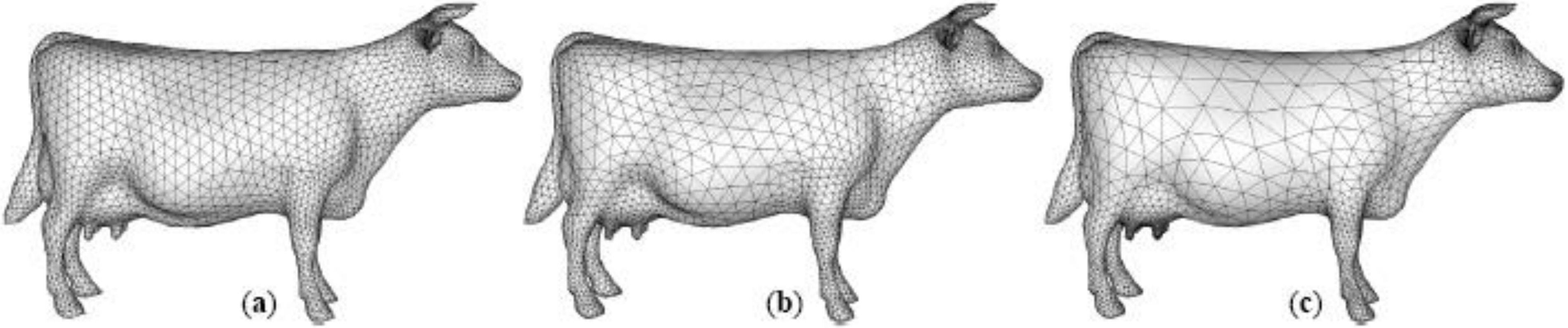
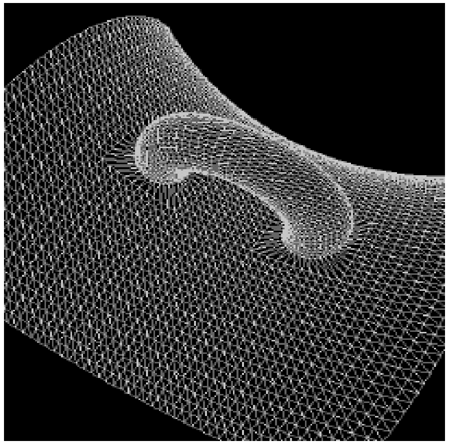
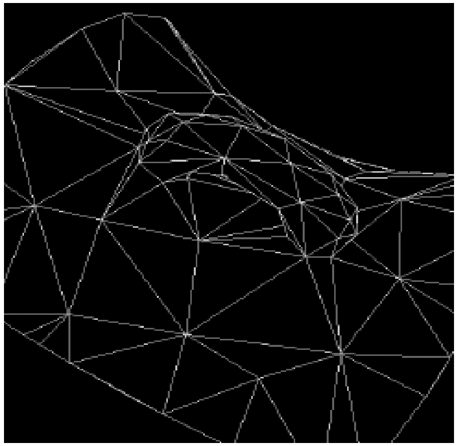
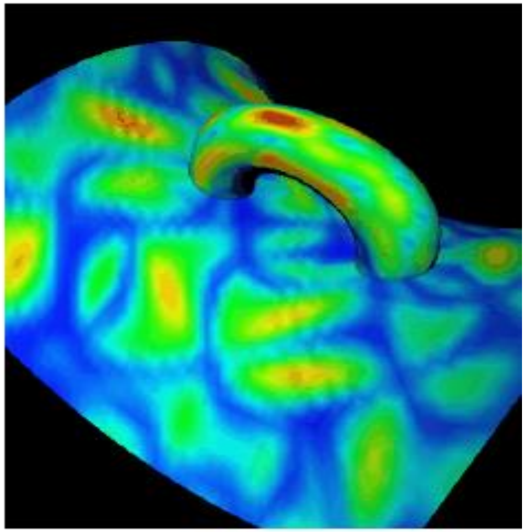
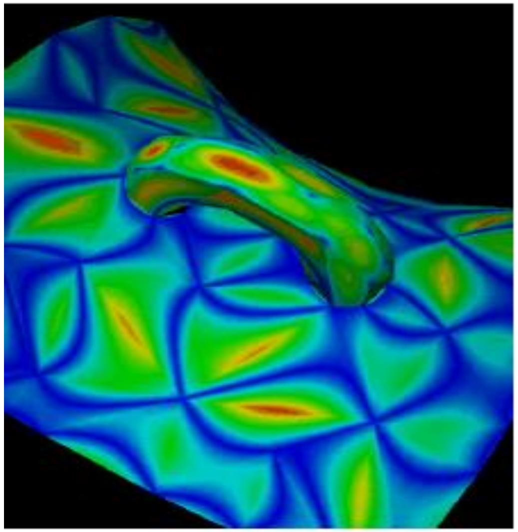


Figure 6: *The cow model with different curvature contrasts.*

Surface Comparison

- METRO (Cignoni_98)
  - Hausdorff Distance for triangulated Meshes

$$E_m(S_1,S_2)=\frac{1}{|S_1|}\int_{S_1} e(p,S_2)\,ds \qquad e(p,S)=\min_{p'\in S} d(p,p')$$

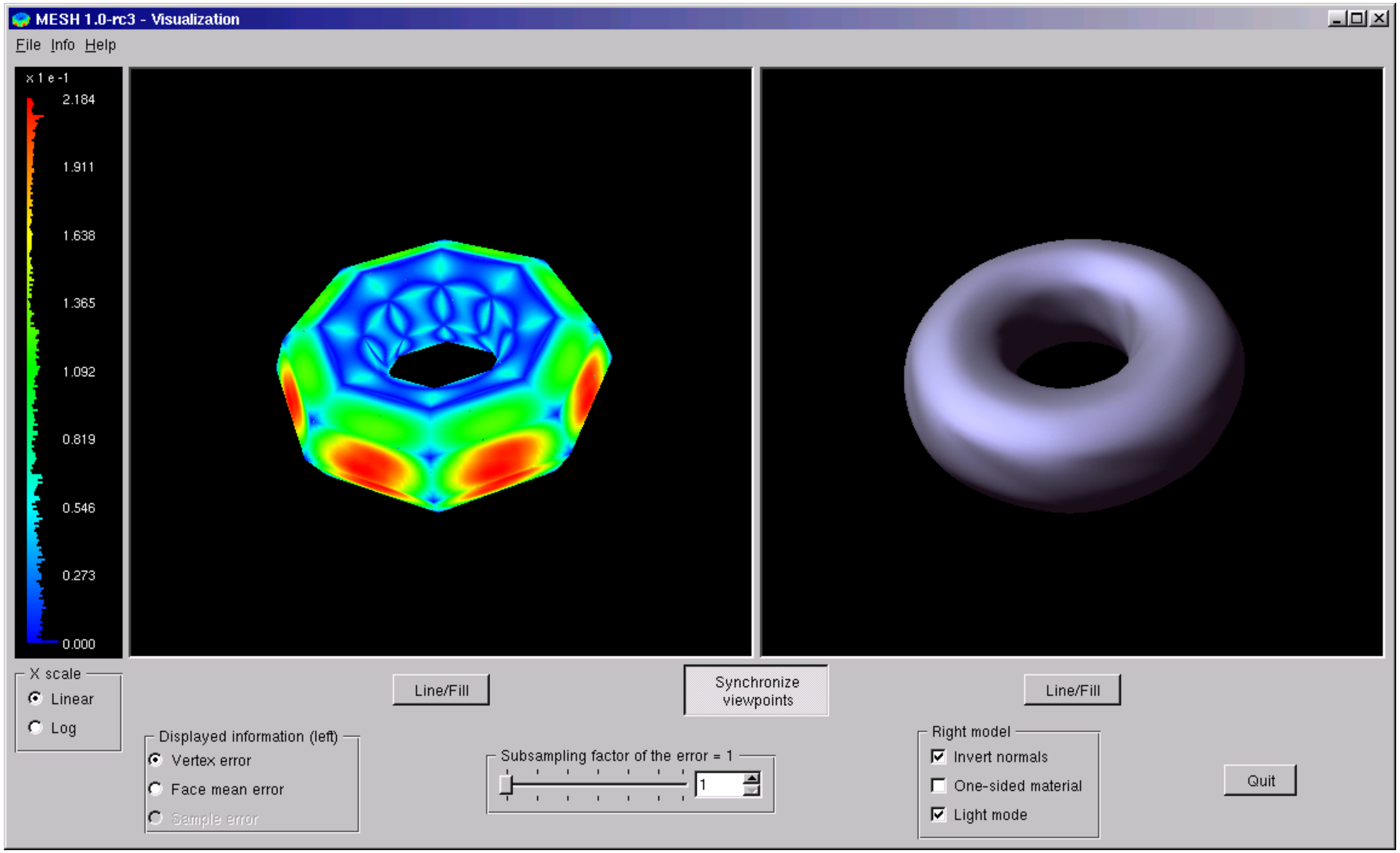
Ouptuts:

- Per-vertex mapping (mean error on the incident faces)

- Error-texture mapping (color-coded error evaluated on each sampling point)

Surface Comparison

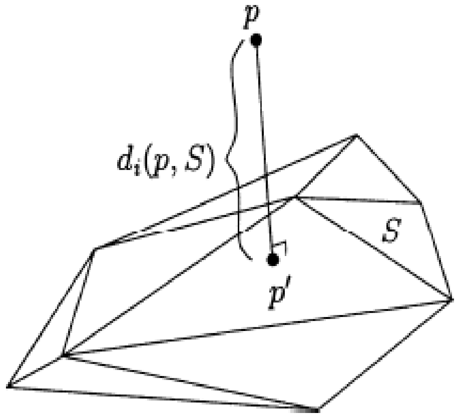
- MESH (Aspert_02)
  - Hausdorff Distance (maximum, mean and mean squared) for triangulated Meshes

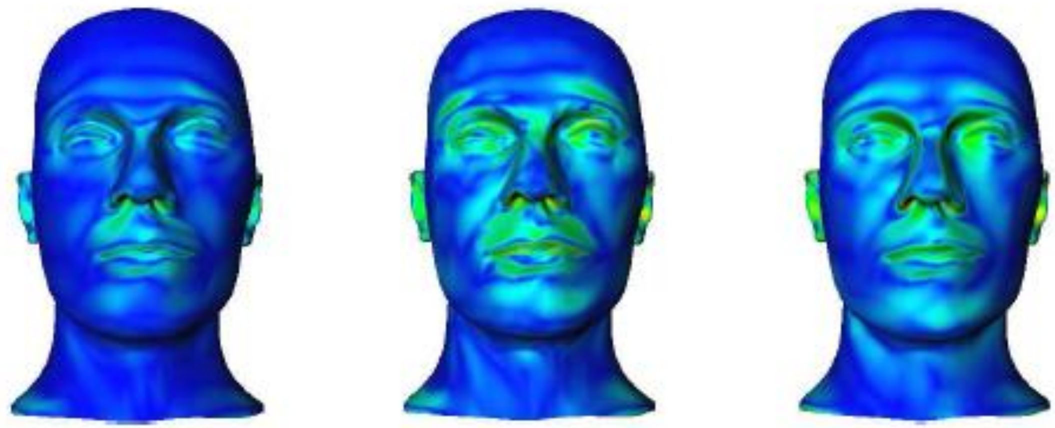


Surface Comparison

- MeshDev (Roy_04)
  - Compare two meshes according to geometrical data or appearance attribute data
  - Uses Attribute Deviation Metric (normal, curvature, ...)

$$d_i(p, S) = \|f_i(p) - f_i(N_S(p))\|$$





(a) QSlim (b) Jade (c) ProgMesh

Fig. 7. Attribute deviation assessment for three different simplification algorithms (the considered attribute is the surface normal).

Topics

- Objectives
- BioClinica Technologies
- Validation in Image Processing
- Evaluation Methodology
- Overview of Evaluation criteria
- Test Data
- Validation Examples
- Conclusion

Test Data

- **Reference / Ground truth / Gold standard**

- Use of realistic simulated images for validation is highly relevant
- As the interest in the computer-aided, quantitative analysis of image data is growing, the need for the validation of such techniques is also increasing.
- Unfortunately, there exists no `ground truth' or gold standard for the analysis of in vivo acquired data.
- One solution to the validation problem is the use of available DATABASES
- Another solution to the validation problem is the use of SIMULATORS, which provide a set of realistic data volumes

Test Data

- Available Databases
  - Internet Brain Segmentation Repository (IBSR) (Medical)
    - http://www.cma.mgh.harvard.edu/ibsr/
  - Berkeley ()
    - http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/
  - GdR-ISIS (Computer-Vision)
    - http://gdr-isis.org/

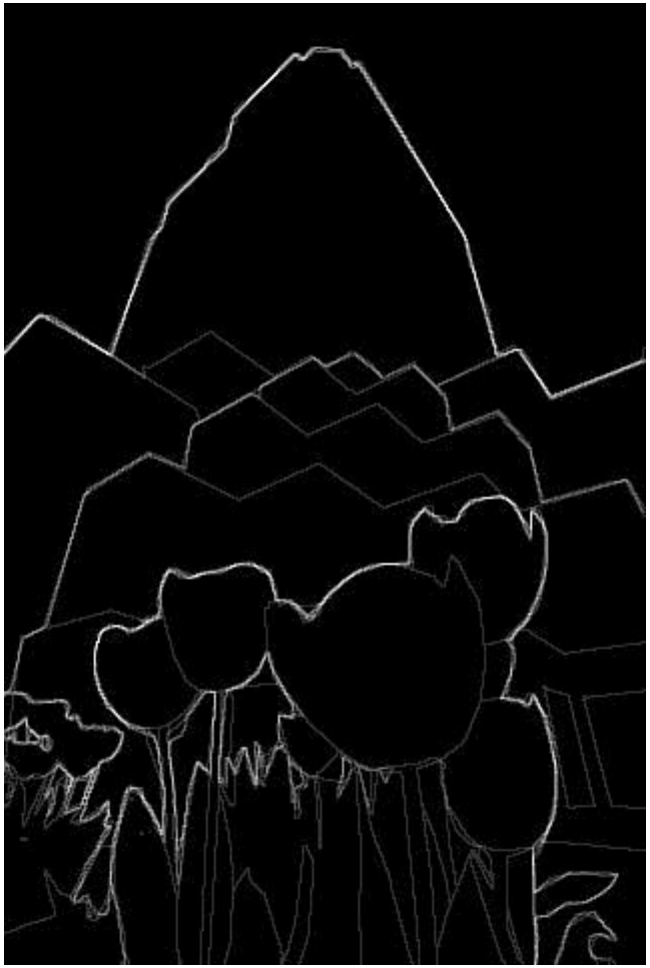
IBSR

- Encourage the development and evaluation of segmentation methods by providing raw test and image data, human expert segmentation results, and methods for comparing segmentation results.
- This repository is meant to contain standard test image data sets which will permit a standardized mechanism for evaluation of the sensitivity of a given analysis method to signal to noise ratio, contrast to noise ratio, shape complexity, degree of partial volume effect, etc. This capability is felt to be essential to further development in the field since many published algorithms tend to only operate successfully under a narrow range of conditions which may not extend to those experienced under the typical clinical imaging setting. This repository is also meant to describe and discuss methods for the comparison of results

Berkeley

- The goal of this work is to provide an empirical basis for research on image segmentation and boundary detection. To this end, they have collected 12,000 hand-labeled segmentations of 1,000 Corel dataset images from 30 human subjects. Half of the segmentations were obtained from presenting the subject with a color image; the other half from presenting a grayscale image. The public benchmark based on this data consists of all of the grayscale and color segmentations for 300 images. The images are divided into a training set of 200 images, and a test set of 100 images.
- They have also generated figure-ground labelings for a subset of these images
- They have used this data for both developing new boundary detection algorithms, and for developing a benchmark for that task. You may download a MATLAB implementation of they boundary detector below, along with code for running the benchmark. They are committed to maintaining a public repository of benchmark results in the spirit of cooperative scientific progress.

Berkeley



GdR-ISIS

- Le GDR 720 ISIS est une structure d'animation du CNRS évaluée tous les 4 ans. De par sa volonté fédératrice, ISIS constitue aujourd'hui un point de passage incontournable et une référence pour la communauté du signal et des images, à laquelle s'est adjointe celle de la vision. Complémentaire d'autres structures d'animation et de diffusion (comme les Colloques GRETSI et RFIA, ou la revue Traitement du Signal), ISIS assure une cohésion nationale à une communauté numériquement importante
- Le GDR est organisé en 4 thèmes placés sous la responsabilité de Directeurs Scientifiques Adjoints :
  - · Thème A - Traitement Statistique de l'Information
  - · Thème B - Image et Vision
  - · Thème C - Adéquation Algorithme-Architecture en traitement du signal et des images
  - · Thème D - Télécommunications : compression, protection, transmission

Test Data

- Available Simulators
  - MRI Simulators
    - BrainWeb (McGill) http://www.bic.mni.mcgill.ca/brainweb/
    - Simri (CREATIS)

    http://www.creatis.insa-lyon.fr/menu/ivolumique/segmentation/simri-hbc/index-us.html
  - PET
    - PET-SORTEO (CERMEP) http://sorteo.cermep.fr/
  - US Simulator
    - Field (http://server.oersted.dtu.dk/personal/jaj/field/)

Test Data

- Conclusion
  - An important effort is done to make databases and simulators available to all the communities
  - Such data will increase efforts in image processing evaluation
  - The use of those available images is useless if you do not consider the way those images were obtained (noise, acquisition, reconstruction ...)

Topics

- Objectives
- BioClinica
- Validation in Image Processing
- Evaluation Methodology
- Overview of Evaluation criteria
- Available Databases / Libraries
- Validation Examples
- Conclusion

Validation Examples

- Illustration on 3 examples

Filtering

- Mean-Shift Filtering
- Correction of Intensity Non-Uniformity Artifact in MRI images

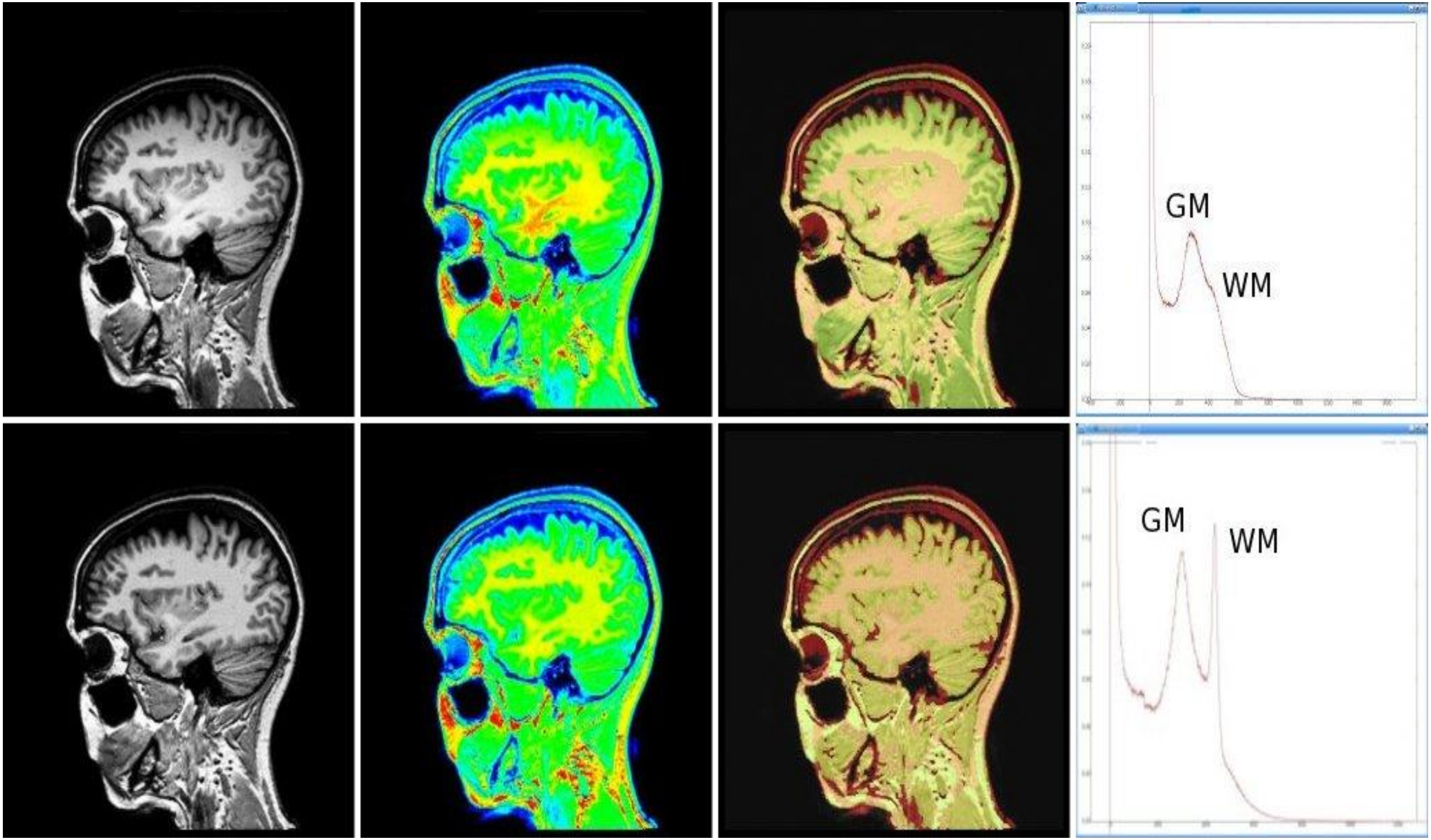
Segmentation

- Impact of Susceptibility Artifact in MRI images segmentation

Correction of Intensity Non-Uniformity Artifact

- Origin
  - RF fields, Acquisition Parameters, Patient interactions
- Impact on MRI Image
  - Low-frequency and smooth intensity variations
- Impact on Image Segmentation
  - Pixels misclassification
- Mathematical Modelling
  - $U(x,y,z) = I(x,y,z).G(x,y,z) + n(x,y,z)$

Correction of Intensity Non-Uniformity Artifact



Validation Examples

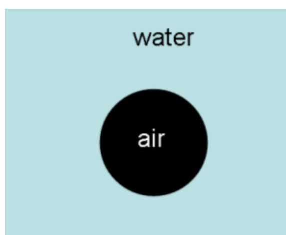
- **Susceptibility Artifact Impact on MRI images Segmentation**

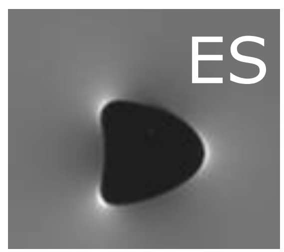
- Origin
  - Magnetic properties of the biological tissues

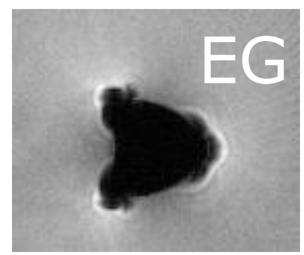
Impact of Susceptibility Artifact

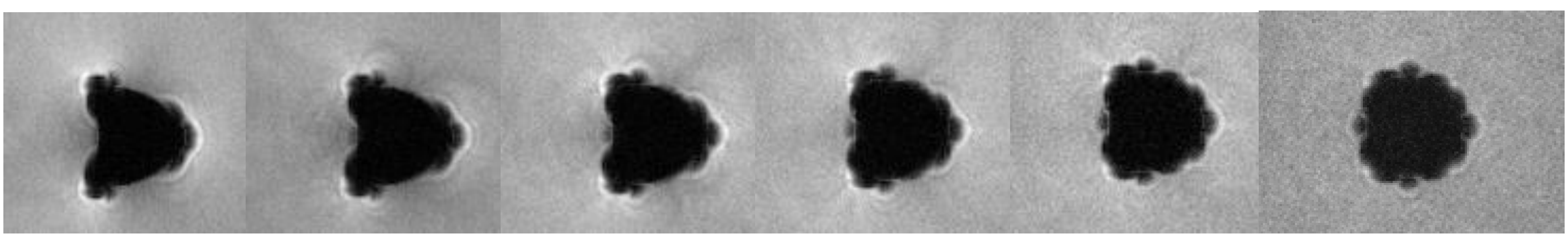
- Impact on MRI images Images
  - Intensity Distortions
  - **Pixel Shifting**

$$\Delta i \propto {}^{1}\!/_{BW}$$









BW = 16 kHz

BW = 200 kHz

Impact of Susceptibility Artifact

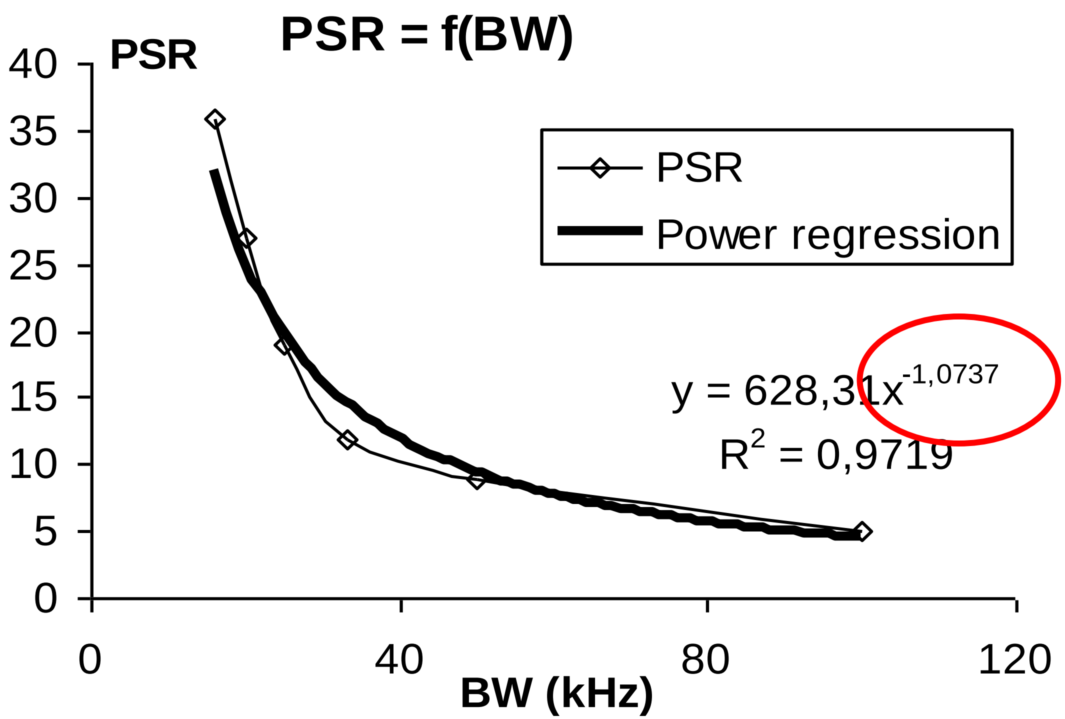
Evaluation Methodology

- Evaluation Criteria

$$EDR = \frac{\sqrt{\sum_{i=1}^{K2} d_o^2(i)}}{\sqrt{\sum_{j=1}^{N} d_A^2(j)}} \qquad IDR = \frac{\sqrt{\sum_{i=1}^{K1} d_u^2(i)}}{\sqrt{\sum_{j=1}^{N} d_A^2(j)}} \qquad PSR = 1 - IDR - EDR$$

Impact of Susceptibility Artifact

- Results on real phantom Data



→ Agreement between theory and experimental results

Topics

- Objectives
- BioClinica
- Validation in Image Processing
- Evaluation Methodology
- Overview of Evaluation criteria
- Available Databases / Libraries
- Validation in clinical trial studies
- Conclusion

Conclusion

- **Some key points about evaluation**

- More efforts have been put on evaluation recently

- However, no many really radical changes / improvements have been widely reported

- Some criteria are deduced from existing ones

- No single evaluation method can be used in all circumstances (Evaluation Guidelines)

- No single evaluation criterion can cover all aspects of segmentation algorithms

Conclusion

- **Limiting factors for evaluation**

- No common mathematical model or general strategy for evaluation

- Difficulties to define wide-ranging performance metrics and statistics

- Testing data used in evaluation are not often representative

- Appropriate gold standard are hard to objectively determine

- Often large costs (time and effort) are involved in performing comprehensive evaluation

Conclusion

Potential research directions

- Efficient Combination of multiple metrics
- Considering the final objective to segmentation to build an evaluation strategy
- Construction of common databases for Image processing evaluation
- Characterize and compare various methods
- Characterize and compare various evaluation methods
- Real use of evaluation results for image processing

Thank you for your attention !