Statistical Analysis of Random Shapes

Vincenzo Capasso, Alessandra Micheletti

Dipartimento di Matematica

Universita' degli Studi di Milano, Italy,

vincenzo.capasso@unimi.it, alessandra.micheletti@unimi.it

In collaboration with:

Andrea Cerri, Patrizio Frosini, Massimo Ferri, Claudia Landi, *University of Bologna*

Silvia Biasotti, Bianca Falcidieno, Daniela Giorgi, *IMATI-CNR Genova*

Gabriel Landini, *School of Dentistry*, *University of Birmingham*

Vittorio Pesce Delfino, Rosalia Ricco, *Policlinico di Bari*

Gabriele Losa, *CERFIM/ISSI Locarno*

Lipari, september 2009

Statistical Shape Analysis

The main aim of Shape Analysis is the description of the mean shape of a random object and the analysis of its stochastic fluctuations.

It is used to face a widespread problem in several applications: *recognize and classify objects and patterns (from digital images) through the description of their shape*.

Mathematical instruments

- Look for mathematical tools allowing…

  - ***to study a finite dimensional problem***
  - ***a "global" description of a shape***
  - ***the analysis of random shapes***
  - ***a good implementation***

LANDMARKS

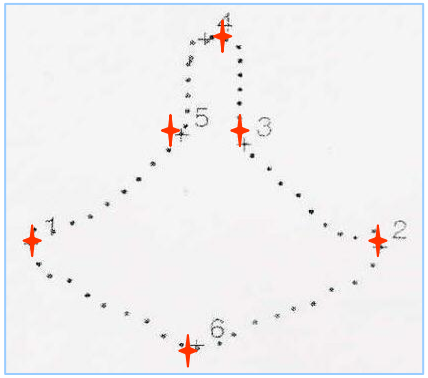
[I.L.Dryden, K.V. Mardia, Statistical Shape Analysis, Wiley, New York, 1998]

PROBLEM I : **REDUCTION OF A GEOMETRIC "OBJECT" BELONGING TO AN INFINITE DIMENSIONAL SPACE (the contour, the colour of internal parts,…) TO AN ELEMENT OF A FINITE DIMENSIONAL SPACE**

LANDMARKS = A FINITE NUMBER OF ELEMENTS (Usually POINTS, but also angles, distances, …) WHICH "CHARACTERIZE" A TYPICAL OBJECT

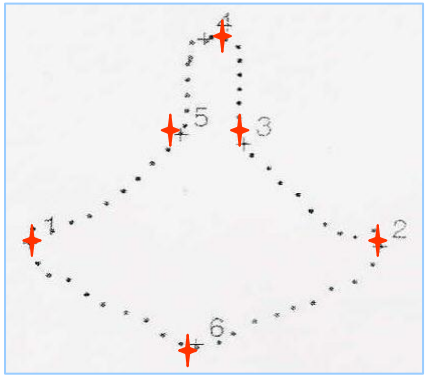
ANATOMICAL LANDMARKS = HAVING A BIOLOGICAL RELEVANCE (ANGLE OF AN EYE, A SPECIAL POINT OF THE SKULL, ETC.)

MATHEMATICAL LANDMARKS = HAVING A MATHEMATICAL RELEVANCE (POINT OF MAXIMUM CURVATURE, ANGLES, DISCONTINUITY POINTS ON THE CONTOUR, ETC.)



EXAMPLE:

6 mathematical landmarks (+) along the contour of a T2 vertebra of a mouse identified as the points of maximum curvature



Every object of the sample is represented by a configuration matrix X, having dimension m x k, containing the m cartesian coordinates of its k landmarks.

If we are interested in the form of the objects, that is if size is relevant, we have to introduce a "summarizing index" also for the size, or **size measure**

A size measure g(X) is a positive function of the configuration matrix such that
g(aX) = ag(X)
for every positive scalar a.

A commonly used size measure is the centroid distance S(X), that is the mean quadratic deviation from the center of mass of the landmarks

$$S(X) = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X}_j)^2}, \quad X \in \mathbf{R}^{km},$$

where X_{ij} is the (i,j)th entry of X and $\bar{X}_j = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$ is the mean over the jth dimension of the landmarks.

Coordinates and Shape Spaces

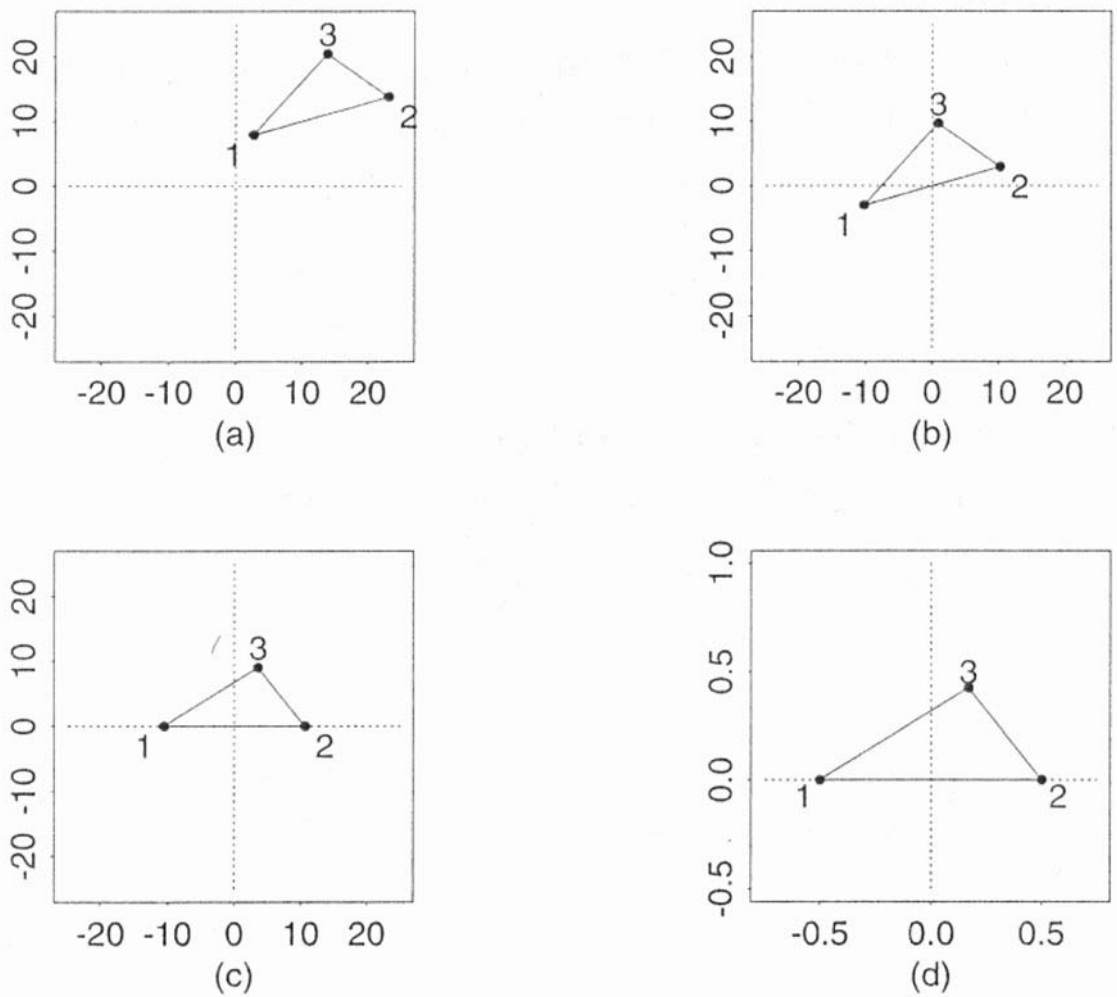
A further dimensional reduction can be obtained by transforming the landmarks coordinates so that the effects of rotations, translations and scaling are eliminated.
This transformation can be performed in various ways, obtaining thus different types of coordinates, and, correspondigly, different types of shape spaces.

First (most straightforward) method: Bookstein Coordinates

Let $(x_j; y_j)$, $(j = 1,\ldots, k)$ be k landmarks in a plane (i.e. we are working in m = 2 dimensions. E.g. on digitized images of objects). Bookstein suggests to label properly the landmarks with numbers and to remove the effects of scaling and rototranslation by shifting, rotating and rescaling all the points so that the landmarks 1 e 2 are moved into the positions (-1/2 , 0), (1/2 , 0), respectively (or, more in general, into two points chosen as symmetric with respect to the origin, to preserve possible symmetries).

The shape coordinates will then be the kx2 pairs of coordinates of the remaining landmarks.

We say that landmarks 1 and 2 form the **basis** of the object.



Transformation of triangle (3 landmarks) in Bookstein coordinates

Kendall Coordinates

Kendall coordinates are similar to Bookstein coordinates, but translations are removed in a different way. To remove translations, in fact, the Helmert submatrix H is used. It is obtained from the k x k Helmert matrix deleting the first line.

The full Helmert matrix H^F of order k, commonly used in Statistics, is a particular orthogonal matrix composed in the following way:

$$H^F = \begin{bmatrix} \frac{1}{\sqrt{k}} & \frac{1}{\sqrt{k}} & \frac{1}{\sqrt{k}} & \frac{1}{\sqrt{k}} & \cdots & \frac{1}{\sqrt{k}} \\ -\frac{1}{\sqrt{2\cdot 1}} & \frac{1}{\sqrt{2\cdot 1}} & 0 & 0 & \cdots & 0 \\ -\frac{1}{\sqrt{3\cdot 2}} & -\frac{1}{\sqrt{3\cdot 2}} & \frac{2}{\sqrt{3\cdot 2}} & 0 & \cdots & 0 \\ \vdots & & & & & \\ -\frac{1}{\sqrt{k(k-1)}} & -\frac{1}{\sqrt{k(k-1)}} & -\frac{1}{\sqrt{k(k-1)}} & -\frac{1}{\sqrt{k(k-1)}} & \cdots & \frac{k-1}{\sqrt{k(k-1)}} \end{bmatrix}.$$

By multiplying X by H^F we obtain a new matrix containing in the first line (corresponding to the first new landmark) the coordinates of the center of mass of the landmarks, multiplied by the factor $1/\sqrt{k}$.

Thus the product of X by H^F corresponds to a translation of the first landmark into the center of mass of the landmarks, eliminating thus the "specific position" of the random object.

The shape variability is described by the coordinates of the remaining k-1 landmarks, that's why the first line of H^F, corresponding to the center of mass of the original landmarks, is neglected.

In general:

In order to define mathematically the concept of *shape,* we need to introduce a suitable metric space so that every shape will be represented by a point in this space.
Also a suitable distance in this space will be introduced which will tell us if and how much two objects will have different shapes.

Distances in the shape space

In order to define the concepts of mean shape and of its variance a distance in the shape space must be introduced, to measure the differences between various objects.

Let us consider two configuration matrices X_1 and X_2 of k points in m dimensions, having preshape Z_1 and Z_2 (*preshape*= configuration after the elimination of translation)

Definition. The full Procrustean distance between X_1 and X_2 is

$$d_F(X_1, X_2) = \inf_{\Gamma \in \text{ rotations},\ \beta \in \text{ scalings}} ||Z_2 - \beta Z_1 \Gamma||$$

Estimate of the mean shape

Let $w_1 \ldots w_n$ be the configurations of the landmarks of *n* centered objects, that is we translate the configurations so that the center of mass of the landmarks of each object coincides with the origin of the coordinate system.

The complete procrustean estimate of the mean shape, or complete procrustean mean shape $[\hat{\mu}]$ is obtained by minimizing with respect to μ the sum of the squares of the complete Procrustean distances of every w_i from a mean unknown configuration μ having unit size, i.e.

$$[\hat{\mu}] = \arg\inf_{\mu} \sum_{i=1}^{n} d_F^2(w_i, \mu).$$

Estimate of the Standard deviation

An estimate of the variability of the shape is provided by the square root of the sum of the deviates of the distances of every configuration from the mean shape, that is the quantity

$$RMS(d_F) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_F^2(w_i, [\hat{\mu}])}.$$

Tangent space coordinates

The tangent space is the linearised version of the shape space (which is NON EUCLIDEAN) in the neighborhood of a particular point (the pole of the tangent projection), which is usually chosen as a mean shape obtained from the dataset of interest. We need to pass from the non-euclidean shape space to a linearised euclidean one in order to apply the usual techniques of multivariate statistics, based on the propagation of gaussian distributions over linear transformations.

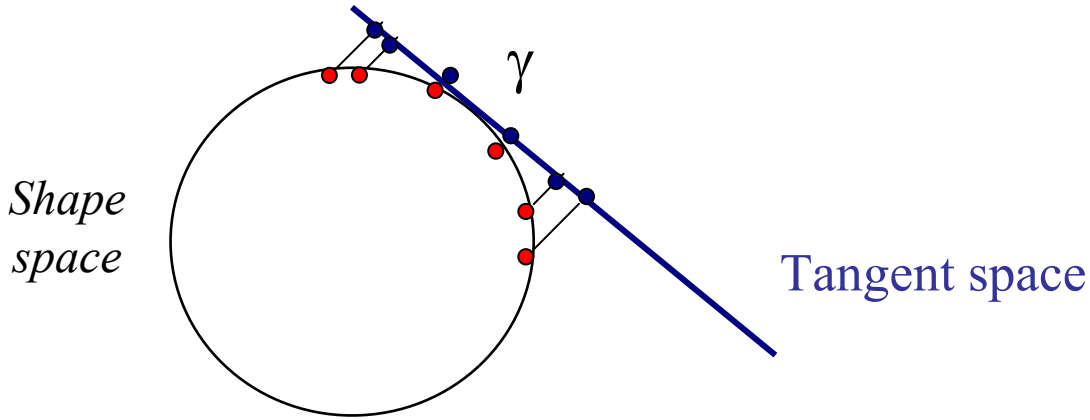
In case of planar data, one of the most used coordinates system is given by the Kent's partial Procrustes tangent coordinates:

Let γ be a complex pole on the complex pre-shape sphere. Let us rotate the configuration by an angle θ to be as close as possible to the pole and then project onto the **tangent plane** at γ, denoted by $T(\gamma)$.

Definition. the **Kent's partial Procrustes tangent coordinates** for a planar shape are given by

$$v = e^{i\hat{\theta}}[I_{k-1} - \gamma\gamma^*]z, \; v \in T(\gamma),$$

where $\hat{\theta} = arg(-\gamma^* z)$.



The approximation via the tangent coordinates is good only if the objects forming the sample are represented by points in the shape space which are CLOSE to the pole γ.

Suitable distances in the tangent space (Euclidean, Mahalanobis, etc.) may lead to automatic classification, via the usual techniques of discriminant analysis.

Hypothesis testing

Hypothesis testing can be performed in the *tangent space*, to compare mean shapes of different samples

$$H_0 : [\mu_1] = [\mu_2] \text{ against } H_1 : [\mu_1] \neq [\mu_2]$$

or to test the equality of the mean shape of a sample with a particular shape, of the type

$$H_0 : [\mu] = [\mu_0] \text{ against } H_1 : [\mu] \neq [\mu_0].$$

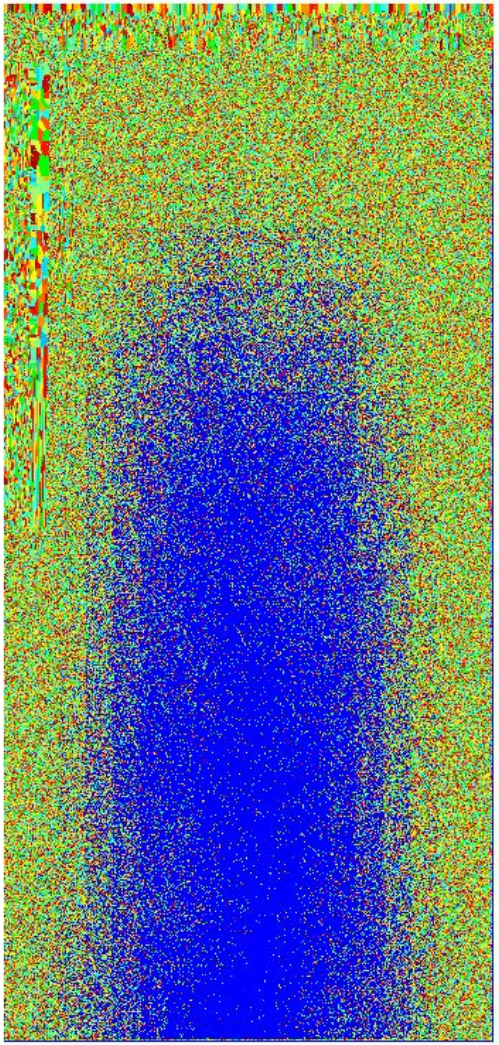
These tests could be used to compare the mean shapes resulting from experiments performed with different physical/chemical/environmental parameters.

Application 1 [Braghieri, Micheletti, 2000]

Temperature dependent polymer crystallization

110 °C

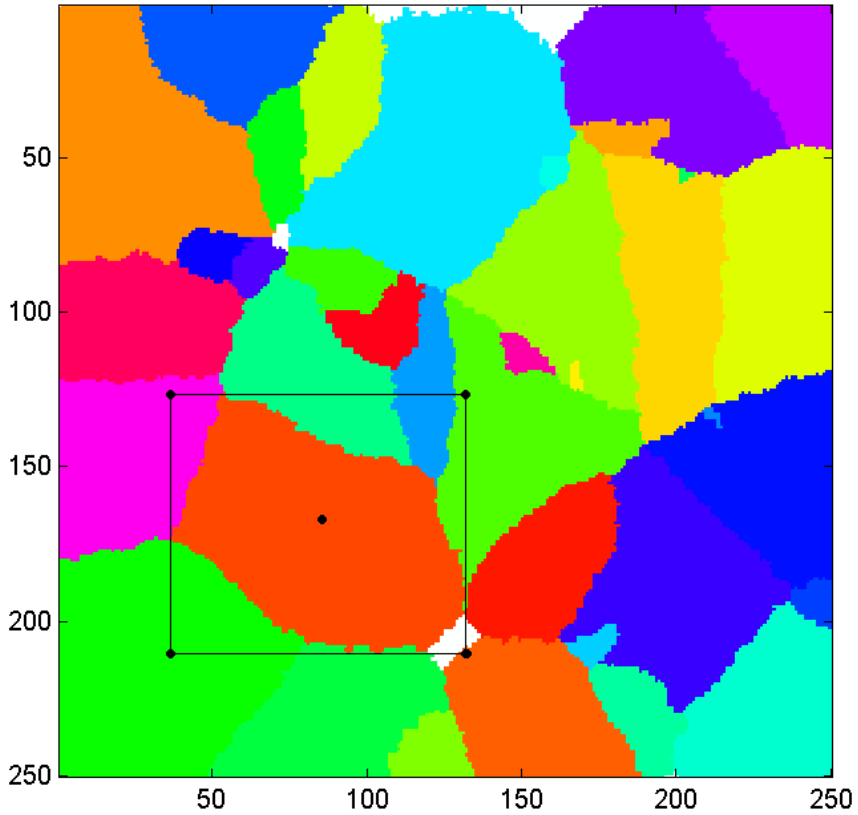
Temperature gradient in the vertical direction



140 °C

What is the minimum temperature difference at which the shape of two crystals is significantly different?

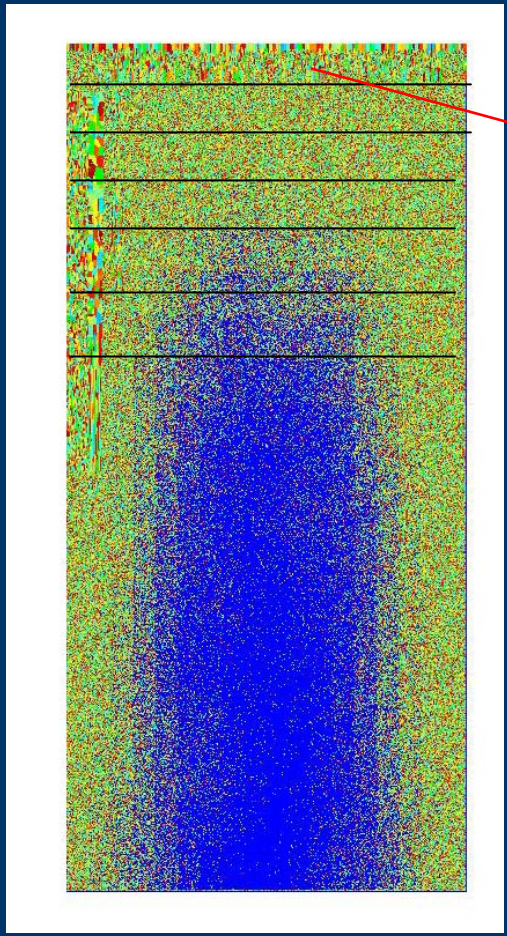
Landmarks:

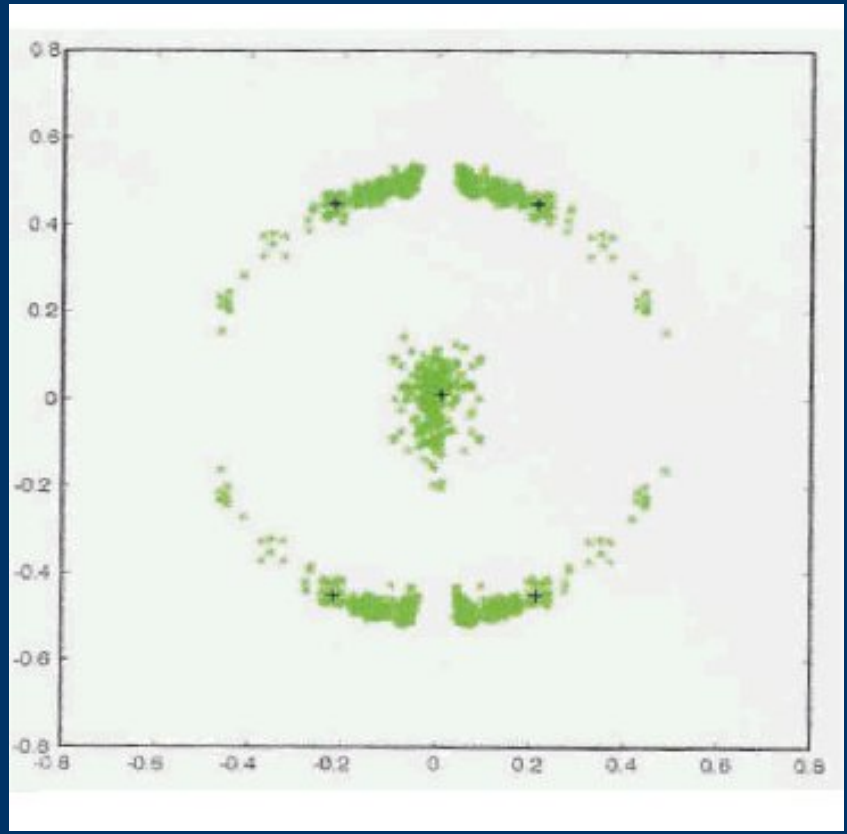


In order to describe the shape of a crystal we used 5 mathematical landmarks chosen as the vertices and the center of the rectangle circumscribed to the crystal

We divided the domain into 6 horizontal stripes, according to the following table. Every crystal belongs to the stripe which contains its center of mass

STRIPE N.	TEMP. OF THE UPPER EDGE	TEMP. OF THE LOWER EDGE	N. OF CRYSTALS
1	110 C	112.5 C	432
2	112.5 C	115 C	129
3	115 C	117.5 C	104
4	117.5 C	120 C	118
5	120 C	122.5	136
6	122.5 C	125 C	26





Clusters of the 5 mathematical landmarks obtained from the sample of simulated crystals coming from the first stripe, using the Kent partial tangent coordinates, and the complete mean procrustean shape (black crosses).

We compared the mean shapes of the different stripes to see if they were significantly different, with the following results (F=value of the statistics, P=correspondent p-value)

stripe	1	2	3	4	5	6
1		F=16.2 P=0.00000	F=23.94 P=0.00000	F=22.16 P=0.00000	F=22.46 P=0.00000	F=2.49 P=0.022
2	F=16.2 P=0.00000		F=3.73 P=0.00147	F=4.8 P=0.00012	F=7.43 P=0.00000	F=4.53 P=0.00029
3	F=23.94 P=0.00000	F=3.73 P=0.00147		F=6.39 P=0.00000	F=4.94 P=0.00008	F=5.35 P=0.00006
4	F=22.16 P=0.00000	F=4.8 P=0.00012	F=6.39 P=0.00000		F=10.95 P=0.00000	F=5.56 P=0.00003
5	F=22.46 P=0.00000	F=7.43 P=0.00000	F=4.94 P=0.00008	F=10.95 P=0.00000		F=8.77 P=0.000001
6	F=2.49 P=0.022	F=4.53 P=0.00029	F=5.35 P=0.00006	F=5.56 P=0.00003	F=8.77 P=0.000001	

Since all the p-values are smaller than 0.05, we may reject H_0 at a 95% level

Problems:

- The choice of Landmarks is crucial, usually is subjective and done by an expert, and different choices may lead to different results.
- In some applications, it is not clear how landmarks can be "ordered" (labeled). Different choices in the labelling may lead to different results.

A need of standards is highly required in this framework!

Size functions + Statistics:

- A statistical-geometrical technique (and related software) for random shapes analysis and classification, with reduction of subjectivity.

We will emphasize two working examples:

1. quality control in litographic processes in microelectronics
2. recognition of normal and tumor cells

Size Functions: definition
[Frosini, Landi, 1999-2004]

Let us consider a pair (M,φ) where

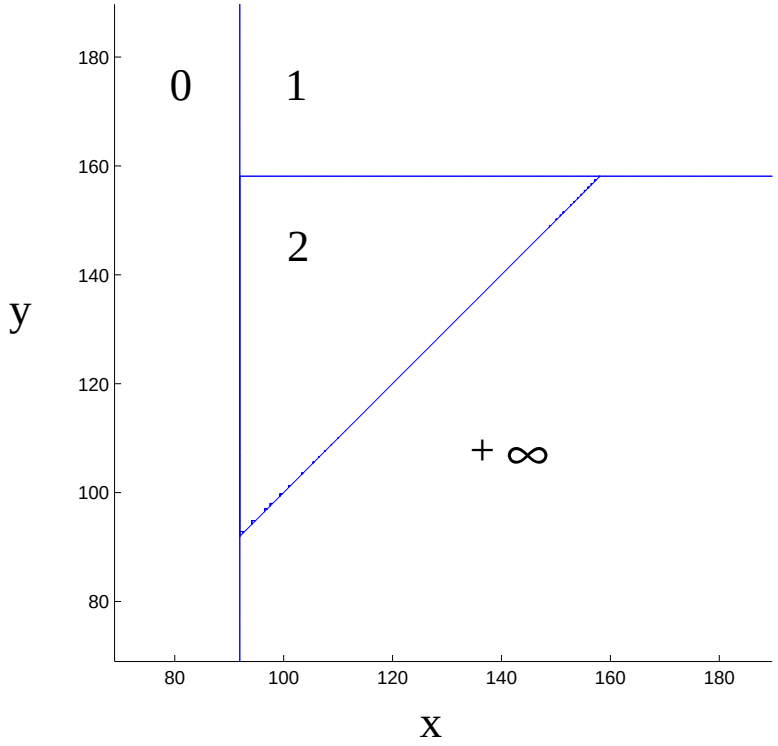
- M is a topological space (sufficiently regular)
- $\varphi : M \to \mathrm{IR}$ is a continuous function, called ***measuring function***

Definition. We shall call *Size Function* associated with (M,φ) the function

$$\mathcal{L}_{(M,\varphi)} : \mathrm{IR} \times \mathrm{IR} \to \mathrm{IN} \cup \{+\infty\}$$

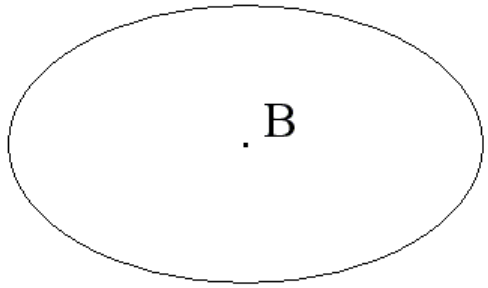
defined by setting $\mathcal{L}_{(M,\varphi)}(x,y)$ equal to the number of equivalence classes into which the set $M\langle\varphi \le x\rangle = \{ \mathrm{P} \in M : \varphi(\mathrm{P}) \le x \}$ is divided by the relation of $\langle\varphi \le y\rangle$ - *homotopy*, where two points $\mathrm{P}, \mathrm{Q} \in M$ are $\langle\varphi \le y\rangle$ - homotopic if and only if either $\mathrm{P} = \mathrm{Q}$ or a continuous path $\gamma : [0,1] \to M$, joining P and Q, exists in M such that $\varphi(\gamma(t)) \le y$ for every $t \in [0,1]$.

Size Functions: example and geometric interpretation



M = *ellipse*

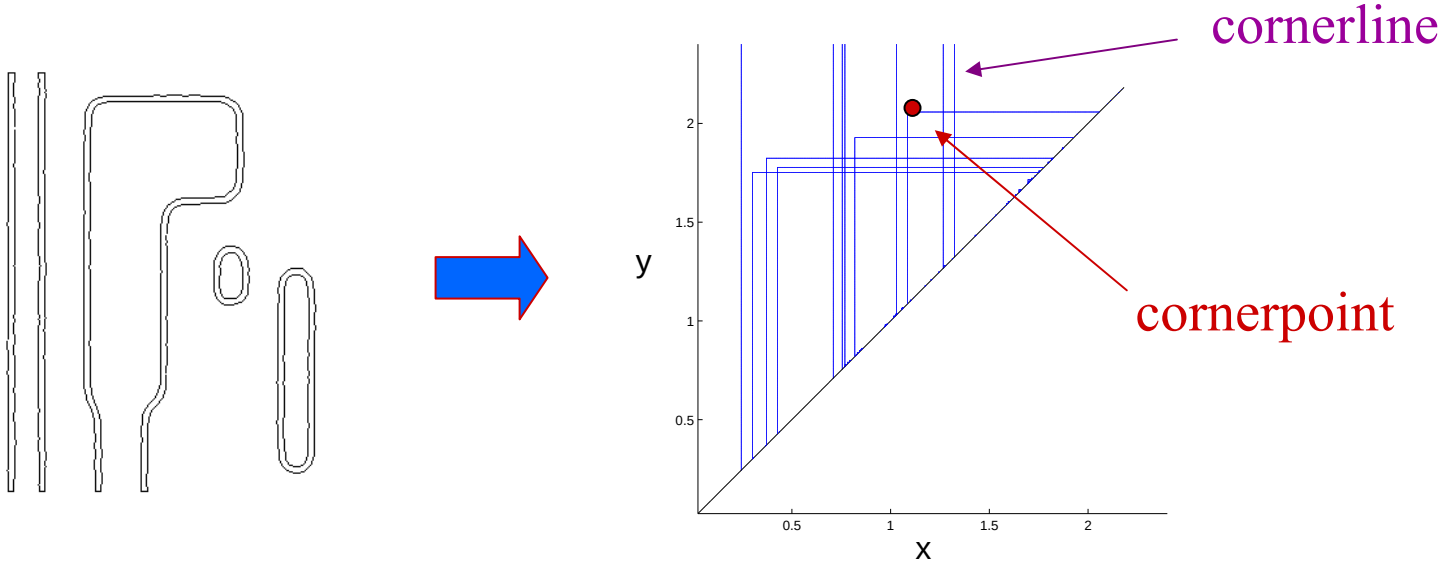
φ = *distance from* B



For $x < y$, $\mathcal{L}_{(M,\varphi)}(x,y)$ *is equal to the number of arcwise connected components of* $M\langle \varphi \leq y \rangle$ *containing at least one point of* $M\langle \varphi \leq x \rangle$.

Size Functions: main properties (1)
[Frosini, Landi, 1999-2004]

- For $x < y$, *cornerpoints* and *cornerlines* uniquely determine the value of $\mathcal{L}(M,\varphi)$ almost everywhere.



all information about the shape under study is conveyed by a finite number of variables.

Size Functions: main properties (2)

- The number of cornerlines (with their multiplicities) is equal to the number of arcwise connected components of the object; the location of cornerpoints is linked to couples of maximum, minimum or saddle values of the measuring function.

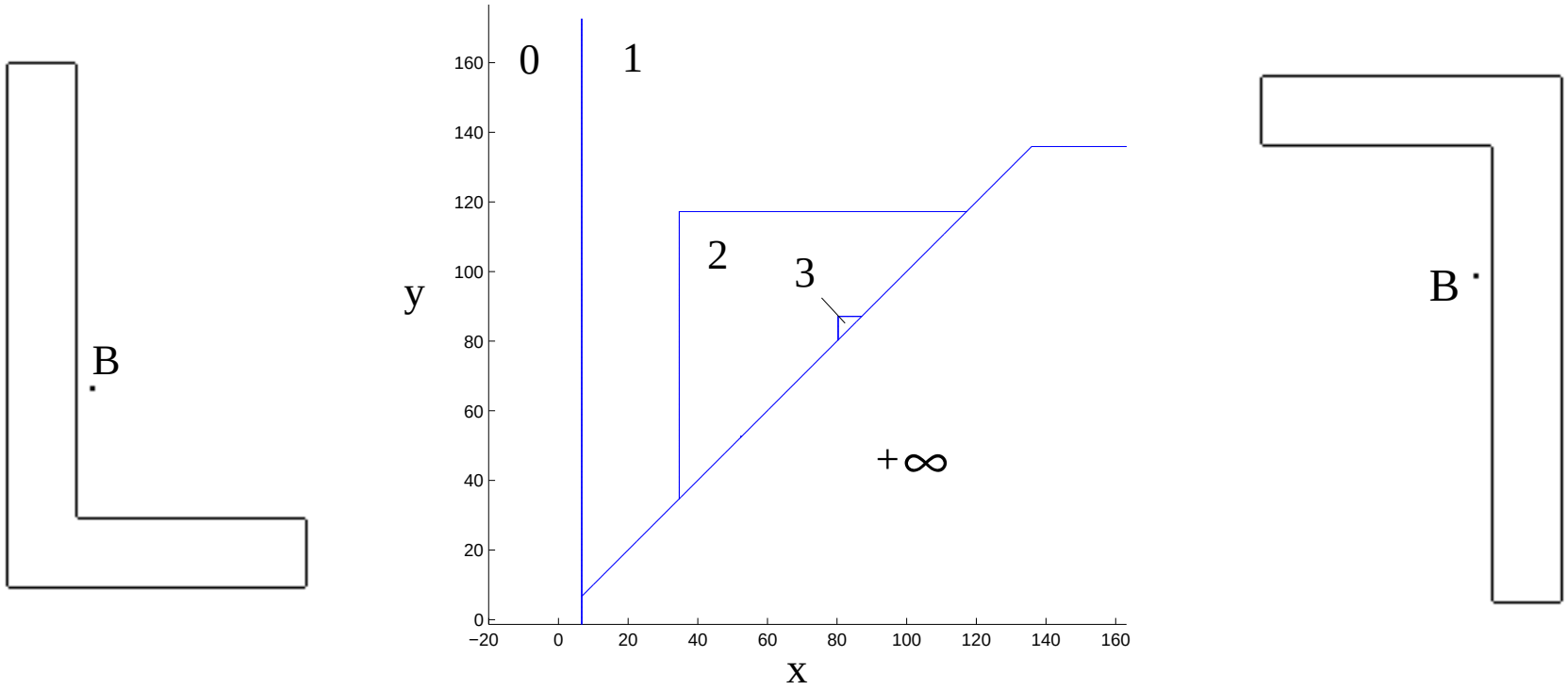
- **The only (crucial!) choice** which must be done for the geometric description of the shape is the **measuring function**.

The crucial choice of the measuring function (1)

- Distinct φ lead to different size functions
- sometimes we need to use more than one φ
- φ gives to the size function *the desired invariance properties*
- a *proper* choice of φ may emphasize *the shape features of the object M which are outstanding* in our application

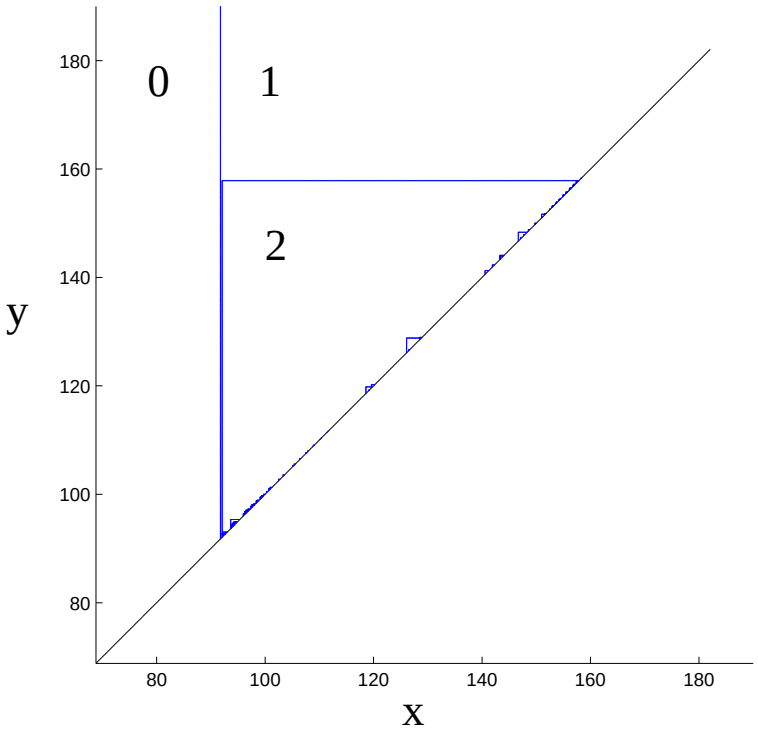
Open problem: choice of the *optimal* φ

The crucial choice of the measuring function (2)



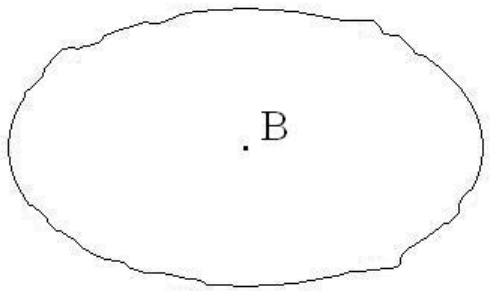
φ = dist. from the center of mass: invariance with respect to isometry

Robustness of the Size Functions with respect to perturbations on a shape



M = *"noisy" ellipse*

φ = *distance from* B

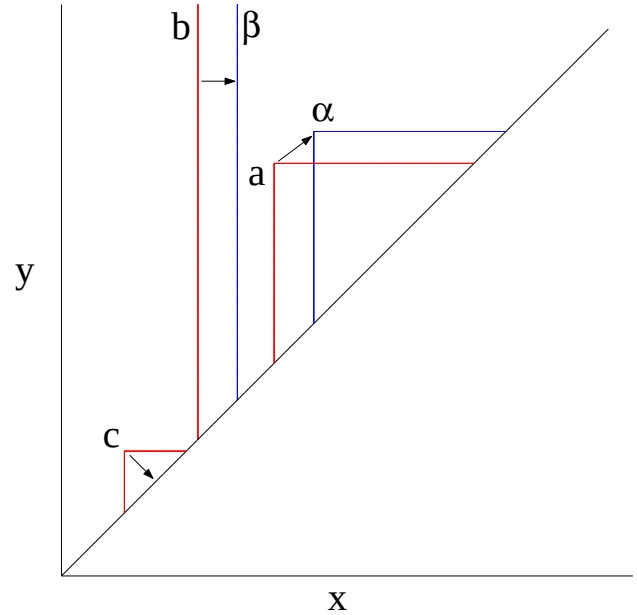


- small displacements of cornerpoints and cornerlines
- small triangles along the diagonal $y = x$

Size Functions comparison: Matching Distance

[D'Amico, Frosini, Landi, 2005]

The *matching distance* compares two size functions by measuring the "cost" of overlapping their sets of cornerpoints and cornerlines, by minimizing the longest movement.



The *cost* to move $p = (x, y)$ to $\bar{p} = (\bar{x}, \bar{y})$ is computed via the *pseudodistance*

$$d(p, \bar{p}) := \min \left\{ \max\{|x - \bar{x}|, |y - \bar{y}|\}, \max \left\{ \frac{y - x}{2}, \frac{\bar{y} - \bar{x}}{2} \right\} \right\}$$

Stability of the Matching Distance

For every $\varepsilon \geq 0$ holds ***[D'Amico, Frosini, Landi, 2005]***

$$\sup_{P \in M} |\varphi(P) - \tilde{\varphi}(P)| \leq \varepsilon \quad \Rightarrow \quad d_{match}(\ell_{(M,\varphi)}, \ell_{(M,\tilde{\varphi})}) \leq \varepsilon$$

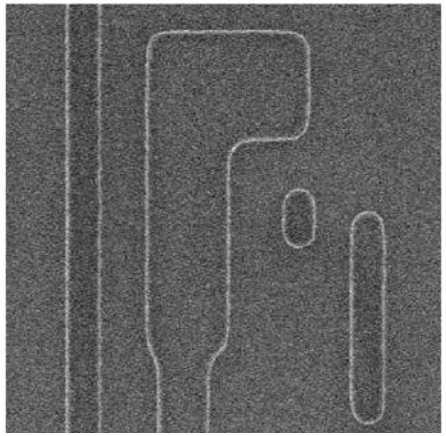
where $\tilde{\varphi}$ is a perturbation of φ due to a perturbation on M.

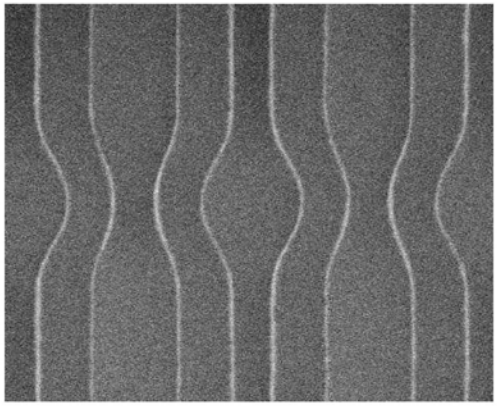
Application 1:litography in microelectronics

[in collaboration with F. Terragni, Universidad Politecnica de Madrid,
M. Vasconi, E. Severgnini,
STMicroelectronics, Agrate Brianza (Milan)]

Analysis of "critical" shapes impressed on a silicon wafer (2)

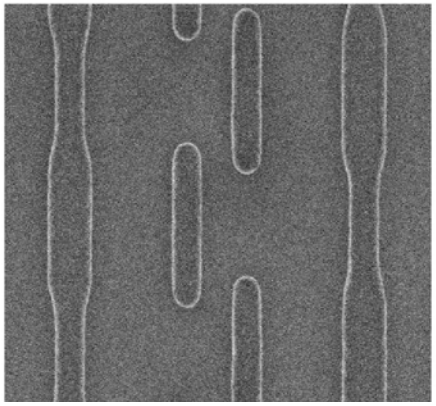
[data collected at STMicroelectronics, Agrate Brianza (Milan)]

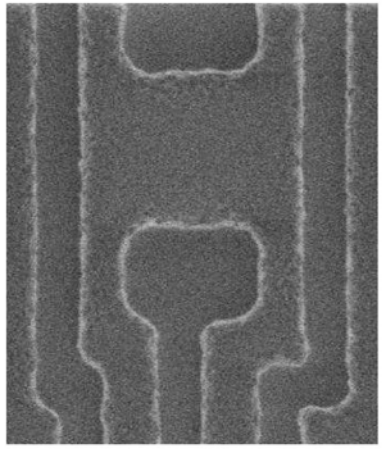




GOAL

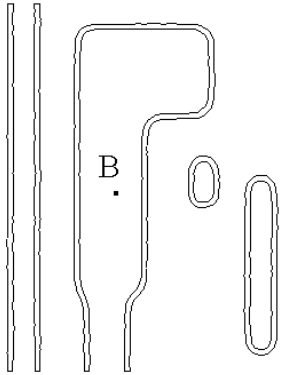
- *"global" description of a sample of shapes impressed with standard process conditions*
- *procedure to classify "good" and "bad" shapes*



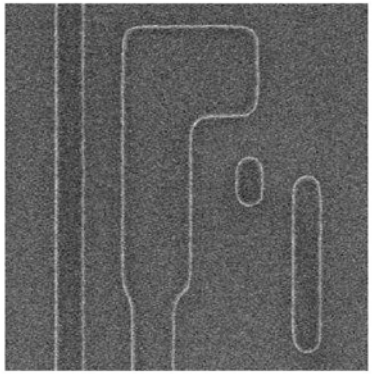


Analysis of "critical" shapes impressed on a silicon wafer (3)

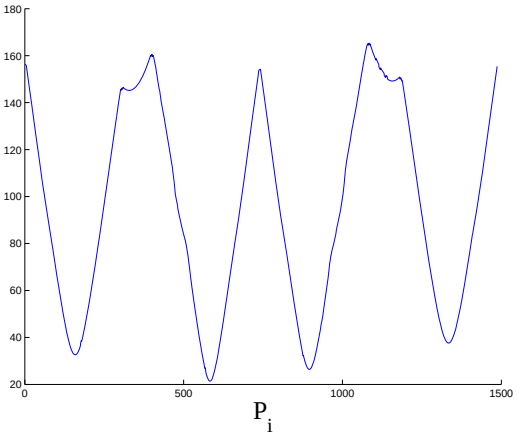
B = *center of mass of the black pixels*



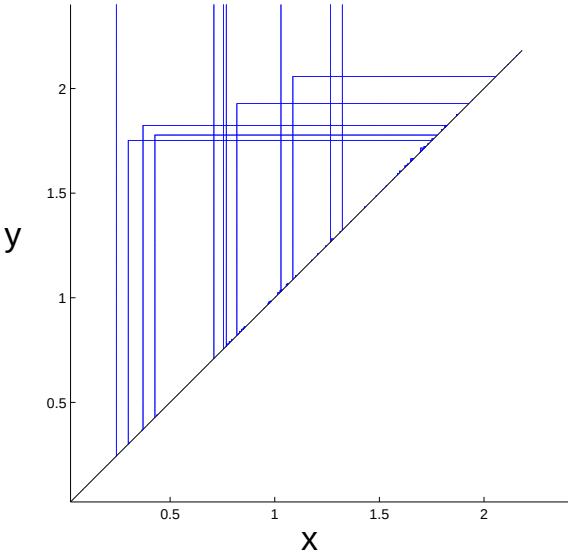
noise filtering
+
edge detection



φ = *dist. from the center of mass*

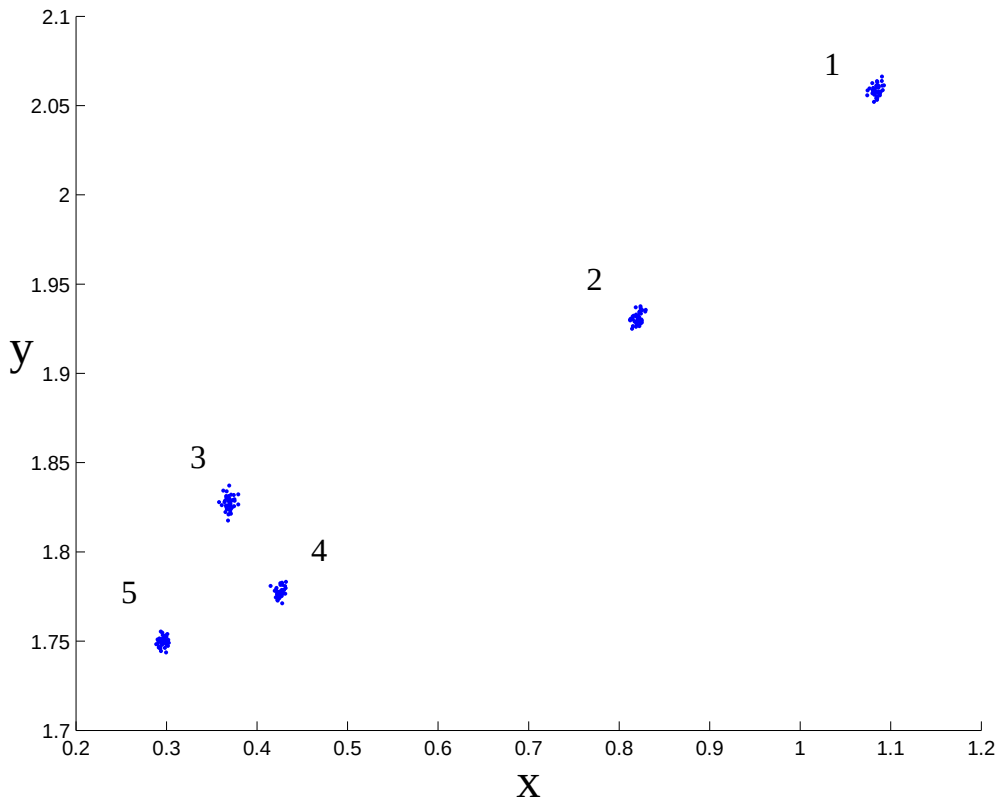


algorithm



Clusters of cornerpoints

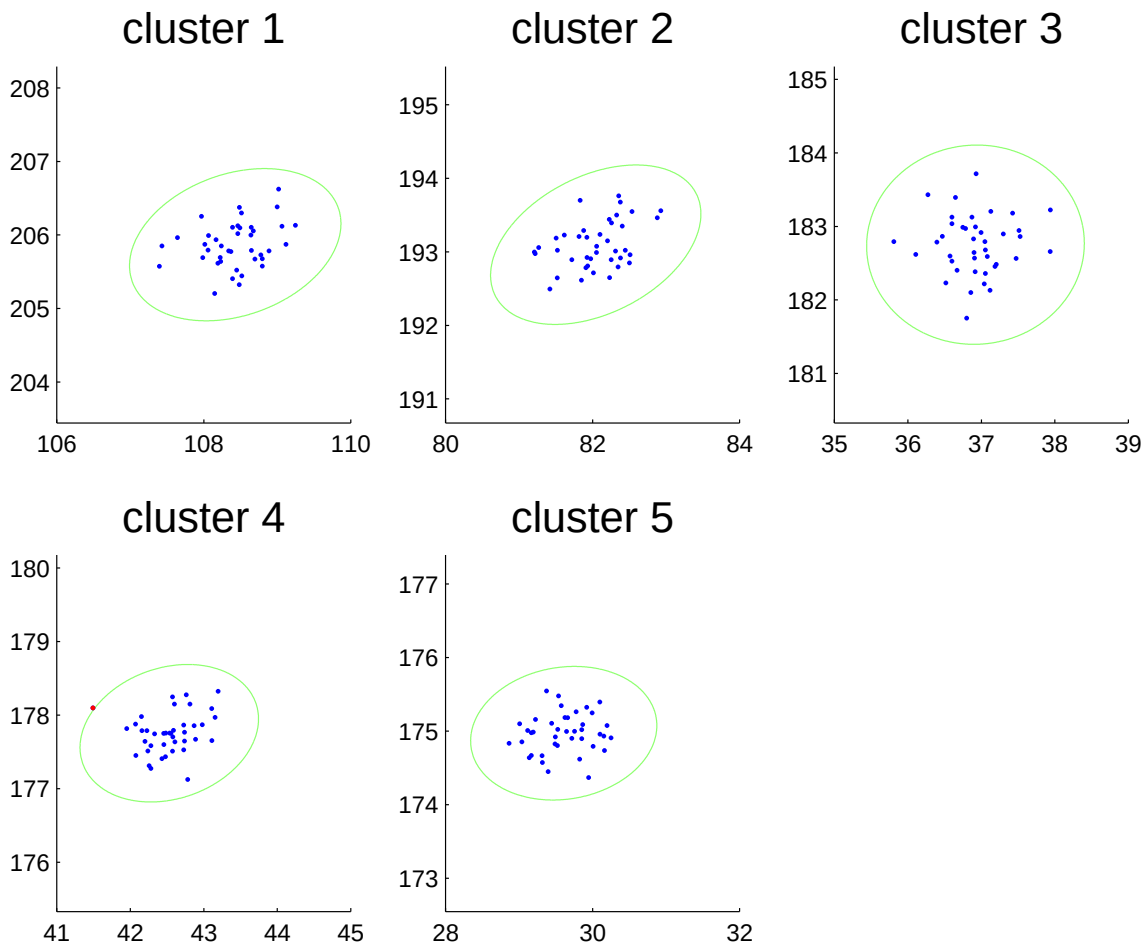
By computing and overlapping the size functions related to different images of the same kind of pattern, we obtain clusters of cornerpoints because of the randomness of its shape and thanks to size functions robustness.
Each cluster is formed by one cornerpoint for each image.



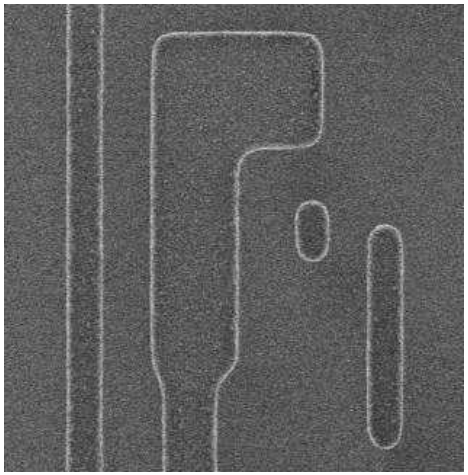
- cornerpoints along $y = x$ due to noise have been filtered out
- cornerlines (identified with their abscissas) can be analysed in a similar way

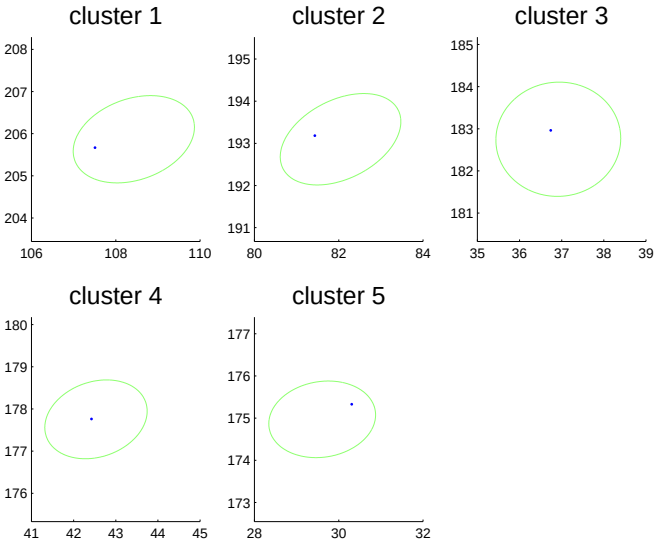
Confidence regions for a "well-impressed" shape

Robustness of size functions
+
Cluster analysis based on the Matching Distance
+
Research of outliers in 2-D samples

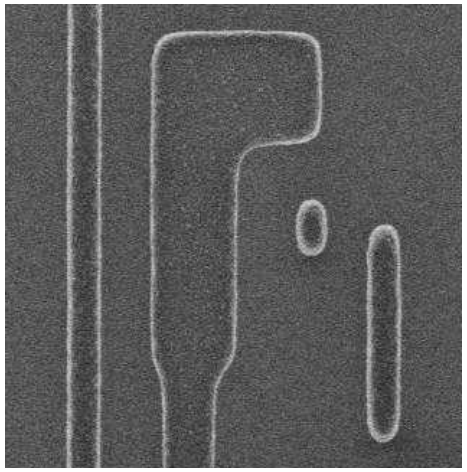


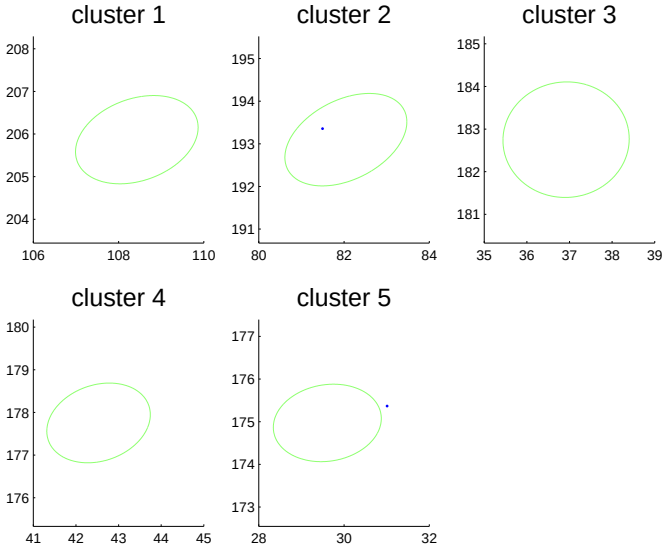
Results





All cornerpoints internal: well-impressed



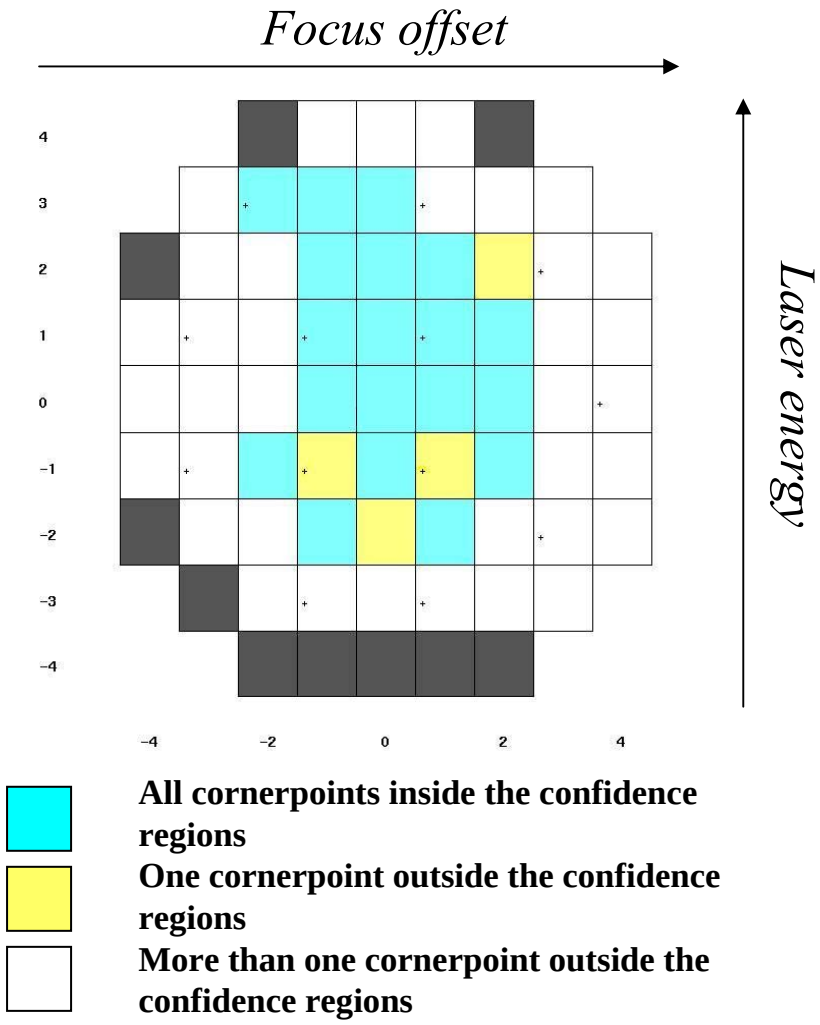


Many cornerpoints external: bad-impressed

Results and Conclusions

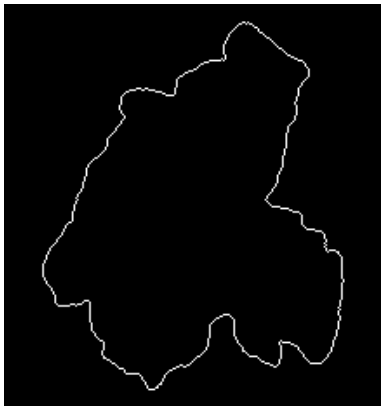
Experiments on a wafer

- Results consistent with expectations.
- Results have evidenced a higher sensitivity of the shapes to the focus offset variation than to the laser energy.
- Global and objective description of the impressed shapes, which can improve the quality control process of devices.
- Automatic procedure to recognize structures not showing "well-impressed" shapes.
- Method which can be extended to other structures or applications.

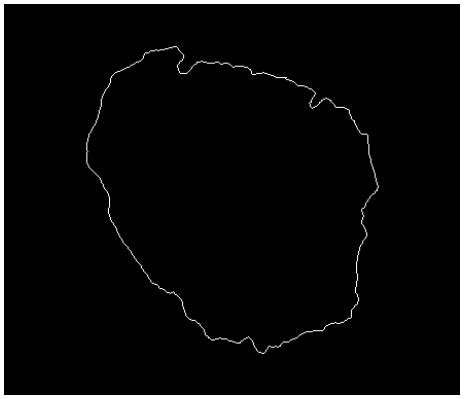


Application 2: automatic recognition of tumor cells

(Data provided by G.Landini, University of Birmigham)



Normal Cell



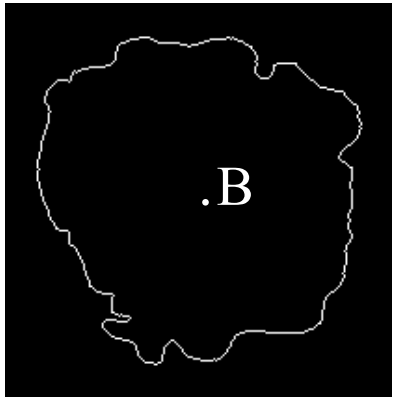
Tumor Cell

Nuclear profiles have been taken from images of cells of istological samples of "normal" or "tumor" zones of the epithelial tissues of the mouth. The nuclei are different mainly because of the different irregularities of the contour.

In order to classify automatically the nuclei, G. Landini used quantities related to the asymptotic fractal dimension of the boundary of the cells to describe their shape (3 landmarks). The Discriminant Analysis applied to these landmarks led to a correct recognition of the 78.8% of the analyzed nuclei (88% of the normal ones and 70.2% of the tumor ones) .

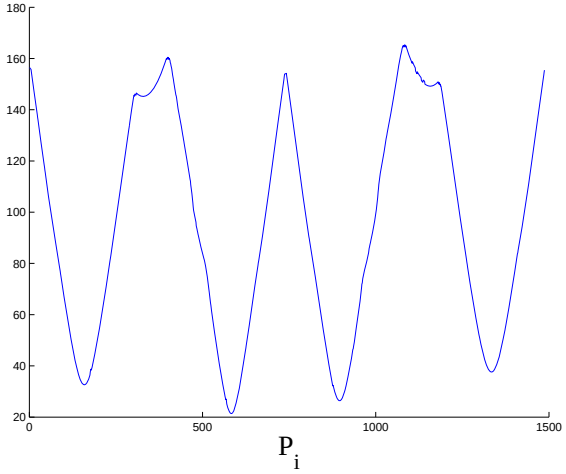
Analysis with size functions

B = *centre of mass of the white pixels*

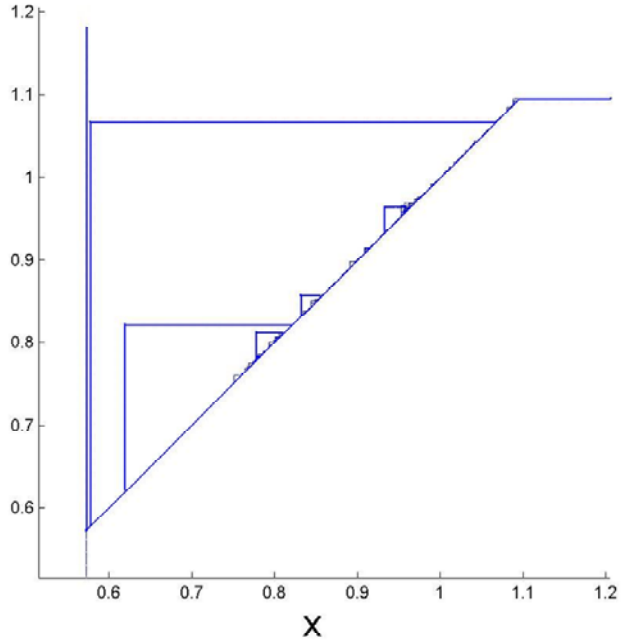


edge detection

φ = *dist. from the centre of mass*

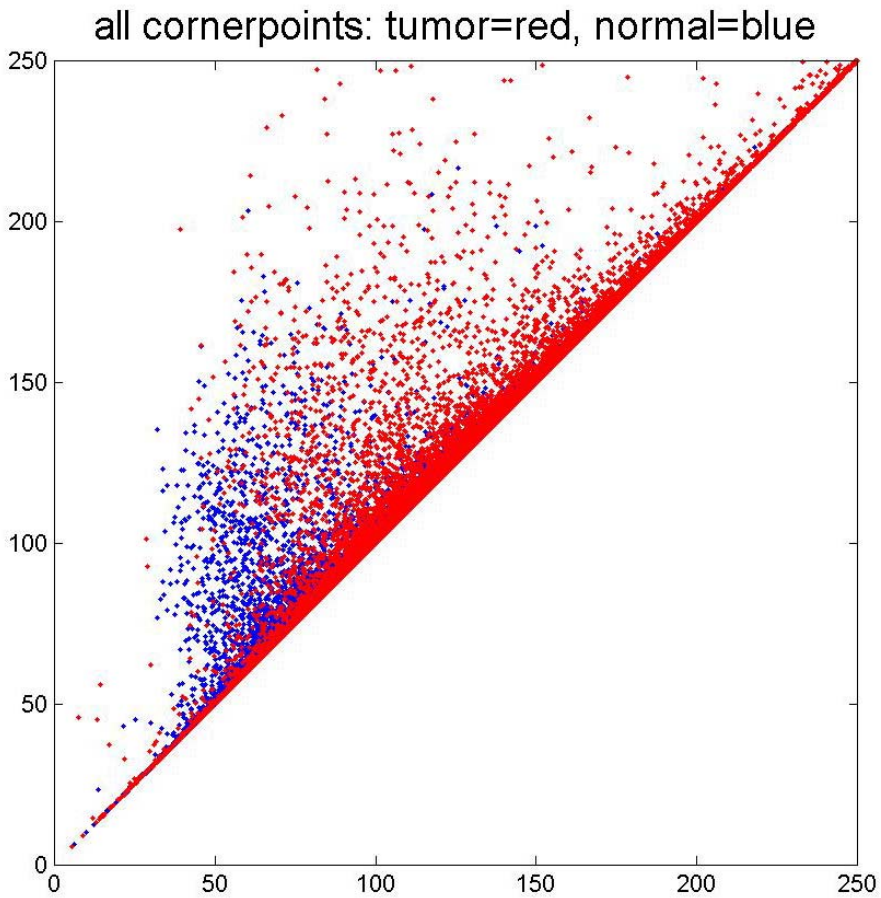


Algorithm (discrete size functions)



Clusters of cornerpoints

By computing the size functions related to different images of the two groups, we obtain two rather mixed clusters of cornerpoints, because of the large variance of the randomness of the shapes.



Classification with NND + matching distance

In order to perform the classification we used a *nearest neighbours distance (NND)* technique:
For any new cell C which must be classified we compute the distance of its size function from the other cells of the sample (already classified), via the matching distance.
Then the *k* cells which are nearest to C are selected, and C is assigned to the class which is more frequent over the *k* nearest neighbours.

k= bandwidth

Classification with NND + matching distance

Problems:

- cornerpoints located close to the diagonal are due to "low level" randomness (small perturbations), which is usually not relevant to identify the class to which the nucleus belongs: we have to filter out the "noise"
- computational costs to compute the matching distance are related to the number *n* of cornerpoints which are retained after the filtering. What is the "correct" filtering?

In our application we retained only the 3 cornerpoints most distant from the diagonal

Experimental Results: φ= distance from center of mass

Sample size= about 700 cells for each class.

% of correct recognition	Of tumor cells	Of normal cells	total
k=7	71.21%	80.79%	75.83%
k=15	70.53%	83.28%	76.68%
k=11	73.84%	80.41%	77.23%
k=17	74.25%	82.02%	78.22%

k= n. of nearest neighbours

k=7 and k=15 are the two possible optimal choices in this case

We used a leave-one-out method in order to compute the percentage of correct recognition

Reminder of Landini's correct recognition:

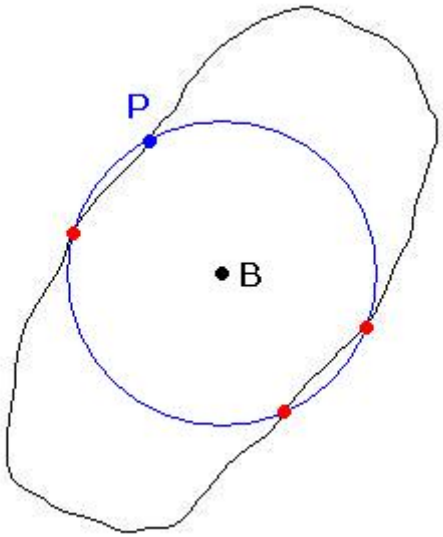
Tumor: 70.2% , Normal: 88% , Total: 78.8%

We also tested the results obtained with another measuring function, defined as follows:

Let B be the center of mass of the image

For every P belonging to the countour of the cell we define $\varphi(P)$ = # of intersections of a circle centered at B passing through P with the contour of the nucleus

Example:



$\varphi(P)=4$

Experimental Results: φ= intersection with a circle

% of correct recognition	Of tumor cells	Of normal cells	total
k=7	35.20%	78.30%	55.97%
k=15	37.11%	76.69%	56.18%

The results are in general worst than what was obtained with the distance from the center of mass

Comparison of the two measuring functions

% of correct recognition k=15	Of tumor cells	Of normal cells	total
Intersections with circle	37.11%	76.69%	56.18%
Distance from center of mass	70.53%	83.28%	76.68%

Using k=15 with both measuring functions, we have
• a better recognition of the normal cells in both cases
• a big improvement in the recognition of tumor cells with the distance from the center of mass

These results are related to the shape of the distributions of the distances of the objects in the two classes: the study of the distributions may lead to a technique to identify an optimal measuring function.

References

- Biasotti, S., Cerri, A., Frosini, P., Giorgi, D., Landi, C.: Multidimensional size functions for shape comparison, *Journal of Math. Imaging and Vision*, 2008.
- Dryden, I.L., Mardia, K.V., Statistical Shape Analysis, Wiley, New York, 1998
- M. D'Amico, P. Frosini, C. Landi, "Natural pseudo-distance and optimal matching between reduced size functions", Preprint.
- P. Frosini, M. Pittore, "New methods for reducing size graphs", *Intern. J. Computer Math.* **70**, pp. 505-517, 1999.
- P. Frosini, C. Landi, "Size theory as a topological tool for computer vision", *Pattern Recognition and Image Analysis* **9**, pp. 596-603, 1999.
- P. Frosini, C. Landi, "Size functions and formal series", *AAECC* **12**, pp. 327-349, 2001.
- A. Micheletti, Statistical shape analysis applied to automatic classification, *Industry Days 2003-2004*, D.Aquilano et al. Eds. The Miriam Project Series, Società Editrice Esculapio, Bologna, pp. 91—100, 2005.
- A. Micheletti, The Theory of Size Functions applied to problems of Statistical Shape Analysis, *Proceedings of S4G,* Prague, 2006.
- A.Micheletti, E. Severgnini, F.Terragni, M.Vasconi, Statistical shape analysis applied to microlithography, *Proceedings of SPIE International Symposium on Microlithography, Data Analysis and Modeling for Patterning Control III,* 2006.
- A.Micheletti, F.Terragni, M.Vasconi, Statistical Aspects of Size Functions for the Description of Random Shapes.Applications to Problems of Lithography in Microelectronics, Progress in Industrial Mathematics at ECMI2006, L.L.Bonilla et al. Eds., Springer - Verlag Heidelberg, pp. 123—134, 2008
- V.Sala, Statistica della Forma applicata a Problemi di Classificazione. Tesi di laurea magistrale, Milano 2008.
- F. Terragni, *Statistical Aspects of Size Functions for the Description of Random Shapes. Applications to Lithography Problems in Microelectronics*, MSc. Thesis, University of Milan, 2005. In Italian.