

## SHAPE AND TEXTURE INDEXES APPLICATION TO CELL NUCLEI CLASSIFICATION

GUILLAUME THIBAUT

*Center for Mathematical Morphology (CMM)  
Mathématiques et Systèmes, Mines-ParisTech  
35 rue St Honoré, Fontainebleau, 77300, France  
thibaulg@ohsu.edu*

BERNARD FERTIL

*LSIS (Laboratoire des Sciences de l'Information et des Systèmes)  
UMR CNRS 6168, Aix-Marseille University  
168 Avenue de Luminy, 13288 Marseille Cedex 9, France*

CLAIRE NAVARRO, SANDRINE PEREIRA,  
PIERRE CAU and NICOLAS LEVY

*INSERM UMR 910, Medical Genetic and Functional Genomic  
Medical School, Marseille, France*

JEAN SEQUEIRA and JEAN-LUC MARI

*LSIS (Laboratoire des Sciences de l'Information et des Systèmes)  
UMR CNRS 6168, Aix-Marseille University  
168 Avenue de Luminy, 13288 Marseille Cedex 9, France*

Received 26 August 2012

Accepted 14 January 2013

Published 25 March 2013

This paper describes the sequence of construction of a cell nuclei classification model by the analysis, the characterization and the classification of shape and texture. We describe first the elaboration of dedicated shape indexes and second the construction of the associated classification submodel. Then we present a new method of texture characterization, based on the construction and the analysis of statistical matrices encoding the texture. The various characterization techniques developed in this paper are systematically compared to previous approaches. In particular, we paid special attention to the results obtained by a versatile classification method using a large range of descriptors dedicated to the characterization of shapes and textures. Finally, the last classifier built with our methods achieved 88% of classification out of the 94% possible.

*Keywords:* Shape and texture indexes; shape and texture characterization; gray level size zone matrix; classification.

## 1. Introduction and Context

Pattern recognition is a major part of artificial intelligence that aims at automating identification of typical situations. It is a major objective for many applications: handwritten character recognition (such as optical character recognition and automatic reading of postal letters and bank checks), video surveillance (facial recognition), medical imaging (ultrasound, CAT scan, Magnetic Resonance Imaging), as well as the context of this paper: computer-aided diagnosis.

At the heart of the pattern recognition, the first and critical step is pattern characterization. Indeed, in order to recognize an object or a person, it is necessary to specifically select the characteristics of interest (morphological, geometrical, textural, ...) and then to identify these characteristics on the digital source under investigation. As a front end in a typical classification system, features extraction is of key significance to the overall system performance. For this reason, it is often helpful to distinguish between two classes of characteristics: the shape using global analysis algorithms<sup>29</sup> (projection histograms,<sup>39,45,55</sup> invariant moments,<sup>13,14</sup> etc.) or outline analysis algorithms (Freeman chain code,<sup>3</sup> MSGPR,<sup>23</sup> Fourier descriptors,<sup>8,45,51,53</sup> etc.) and the texture<sup>41,46,54</sup> using statistical analysis,<sup>9,16,17</sup> wavelets transform<sup>8,47,48</sup> or local binary pattern<sup>33,34</sup> for example.

The aim of this paper is to create a model to classify cultured skin fibroblast nuclei in patients affected by Progeria disease (also known as Hutchinson–Gilford-Progeria syndrome). In 2003, a major research advance<sup>11,37</sup> showed that this rare disease (which affects about 100 patients in the world) is caused by a mutation in LMNA gene on Chromosome 1. LMNA gene codes for lamins A and C, two proteins localized at the nuclear periphery and within nucleoplasm. Progeria patients exhibit an accelerated aging. The presence of mutated lamin A protein results in abnormal nuclear shape and texture, as revealed by the immunodetection of lamin A/C (primary antibody directed against lamin A/C, secondary antibody coupled to fluoresceine isothiocyanate (FITC), see Fig. 1). The digitized pictures of immunostained nuclei were sampled using a conventional confocal bi-photonics epifluorescent microscope (Leica DMR) coupled to a Princeton-Roper camera.

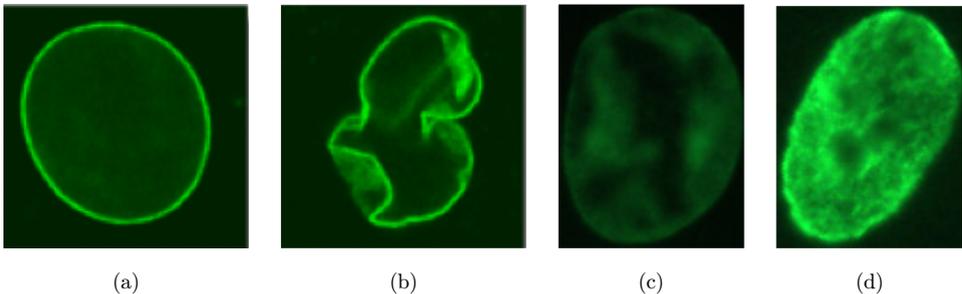


Fig. 1. Examples of nuclei highlighted with FITC: (a) healthy, (b) puffy shape, (c) and (d) heterogeneous texture. Digitized pictures of immunostained nuclei were sampled using a conventional epifluorescent microscope (Leica DMR) coupled to a Princeton-Roper camera.

According to the expert's experience, the two main clues for the presence of the mutated gene are the shape [*normal* or *puffy*, see Figs. 1(a) and 1(b)] and the texture (*homogeneous* or *heterogeneous*, see Figs. 1(c) and 1(d)). But at present time, nuclei labeling is highly time consuming: it requires at least one day to an expert to label approximately 300 nuclei, so a small sample according to the nuclei is acquired. Moreover, it appears that an expert does not provide the equal labels to difficult cases when he labels twice the same sample (ratio of repeatability lower than 1). So the two goals of this study are:

- to design an automatic classification method of abnormal nuclei matching microscopist expert efficiency, thus enabling the analysis of large numbers of nuclei with the same accuracy;
- to use this automated procedure to follow up the eventual nuclear changes induced by therapeutic drugs.

## 2. Classification

The aim of classification is to attribute a class to each object under consideration. In our context, it should be a two-class problem: the objective is to determine whether a cell nucleus is normal (*healthy*) or abnormal (*pathological*). However, we benefit from the knowledge of biologists and geneticists who not only have specified classes (*healthy* and *pathological*), but also specified subclasses (*normal* and *irregular* shapes, *homogeneous* and *nonhomogeneous* textures, which are the two main clues in the presence of the mutated gene). We subsequently have identified a series of relatively independent subproblems that may be solved independently. In addition, since a set of 2800 fully expertised nuclei is available, we can therefore take advantage of supervised methods to design the various classifiers required for this study.

Such classifiers are usually built using a learning method, with the help of cross-validation to warrant generalization (i.e. their efficiency must not degrade for new data). According to the cross-validation protocol, the data are separated into two groups: a learning set and a validation set. The classifiers are built using the learning set but are expected to reach the same level of performance on the validation set.

Prior to the classification phase, it is necessary to construct a vector of characteristics describing the data. The vector must be relevant to the problem in order to allow accurate classification and prediction. The major risk when providing too many characteristics to the classifier is *overfitting*. The greater the dimension of the characteristic vector, the greater the flexibility of the model and the better the classification on the learning set, but the greater the likelihood that the model's performance will be poor for new data. Each model must then be systematically validated and the best one for the validation set is selected. In this paper, the validation is done according to *K-Fold Cross Validation*<sup>10,40</sup> or *leave one out* (LOO) protocol<sup>10,30</sup> (a *K*-fold validation with *k* equal to the size of the working set). The choice of protocol depends on the size the working set.

The main model chosen for classification is the logistic regression.<sup>2,19,32</sup> It is a linear model particularly convenient for two-class classification problems:  $P = P(Y/\vec{x}) = \frac{e^{f(\vec{x})}}{1+e^{f(\vec{x})}}$  with  $\vec{x} = (x_1, \dots, x_n)$  the characteristic vector of the input data,  $f(\vec{x}) = \sum_i \alpha_i x_i$  and  $P(Y/\vec{x})$  the conditional probability  $P$  of the variable  $\vec{x}$  to belong to the class  $Y$ . The likelihood function  $L(\alpha, Y) = P^Y [1 - P]^{1-Y}$  is used to estimate the coefficients  $\alpha_i$  of the model. Nonlinear models are also used in order to analyze results with more details:

- *K-nearest neighbors*, one of the oldest, simplest and most intuitive nonlinear methods of classification.<sup>12</sup> We have systematically tested various numbers of clusters ( $k$ ).
- *Random forests* (RF),<sup>6</sup> a nonlinear classification technique based on the use of *classification and regression trees* (CART)<sup>4</sup> for the building of each tree. It is one of the most recent developments in the research of the aggregation of randomized decision trees. It synthesizes approaches developed in Refs. 1 and 5.
- *Neural networks*,<sup>31</sup> a method widely used in classification, thanks to its potential of modeling (it can approximate any sufficiently regular function). It can solve a large variety of problems, including complex phenomena, complex data that eventually do not follow probabilistic laws. We use a *Multi-Layer Perceptron* (MLP)<sup>36</sup> with one hidden layer. The number of nodes in the hidden layer was calculated according to the number of neurons in the input and output layers:  $(N_{\text{Input}} + N_{\text{Output}})/\nu$ . We systematically tested various values of the parameter  $\nu$  in order to find the best network design.

**Remark** about the dataset of nuclei. A set of 2800 nuclei labeled by experts is available, but only 135 of the nuclei (5% of the dataset) have a nonhomogeneous texture. As far as texture is concerned, building a classifier may be tricky since the sample size of the two classes is particularly unbalanced. In that case, we implemented an under-sampling protocol<sup>25</sup> in order to give the two classes comparable sizes: the majority class (nuclei with homogeneous textures) is reduced to a selection of the 135 most representative nuclei (paragons). The paragons are selected by a 135-*means* clustering,<sup>18</sup> which is turned robust<sup>a</sup> by the determination of strong forms as observed after several runs of the algorithm. Paragons are the nearest nuclei of the barycenters of strong forms. A working set of 270 nuclei (135 paragons and the 135 nuclei with nonhomogeneous texture) is subsequently obtained. Note that other techniques are available to deal with unbalanced data. It is possible to implement *over-sampling*<sup>26</sup>: random or directed duplication of nuclei of the minority class (nonhomogeneous) until the sets are balanced. However, it is shown in Refs. 21, 22 and 25 that this technique is *generally* less efficient than under-sampling. Other techniques based on asymmetric entropy measure<sup>28</sup> or the use of an auto-associator neural network<sup>21</sup> may also be used as well.

<sup>a</sup> *K*-means properties that make consecutive iterations provide different results. But the strong forms are groups of nuclei systematically together, so paragons are highly stabilized results.

Finally, the resulting small size of the working set for texture classification (270 nuclei) suggests validation by the Loo protocol.<sup>30</sup>

### 3. Shape Characterization and Classification

Shape analysis is the most important issue for the diagnosis of nuclei<sup>24</sup>: the study of expert's annotations reveals that 80% of pathological nuclei have abnormal (*puffy*) shapes.

#### 3.1. Notations

Let  $f(\mathbf{x}) : \mathcal{E} \rightarrow \mathcal{T}$  be a gray-level image, where  $\mathcal{E} \subset \mathbb{Z}^2$  is the space pixels  $\mathbf{x} = (x, y) \in \mathcal{E}$  and the image intensities are discrete values which range in a closed set  $\mathcal{T} = \{t_1, t_2, \dots, t_N\}$ ,  $\Delta t = t_{i+1} - t_i$ , e.g. for a eight-bit image we have  $t_1 = 1$ ,  $N = 256$  and  $\Delta t = 1$ .

Let assume also that image  $f$  is segmented into its  $J$  flat zones (i.e. connected regions of constant value, see Sec. 4):  $\mathcal{E} = \cup_{j=1}^J R_j[f]$ ,  $\cap_{j=1}^J R_j[f] = \emptyset$ . The size (surface area) of each region is  $s(j) = |R_j[f]|$ . Hence, we consider that each zone  $R_j[f]$  has a constant gray-level intensity  $g(j)$ .

Moreover, we note  $\varphi \in f$  a connex binary pattern included in  $f$ , with  $A(\varphi) = |\varphi|$  the surface (noted  $A$  if unambiguous) and  $P$  the perimeter (notations are given at the end of the paper).

#### 3.2. Previous works

##### 3.2.1. Moments: Hu, Legendre and Zernike

*Moments* in mathematics (notably in probability) are derived from moments in physics. The moment  $m_n(\lambda)$  of order  $n$  of a function  $\lambda$  is defined on an interval  $\Upsilon$  (not reduced to a point) of  $\mathbb{R}$ .

$$m_n(\lambda) = \int_{\Upsilon} x^n \lambda(x) dx.$$

So the centered moment of order  $n = p + q$  applied to an image with a pattern  $\varphi$  of centroid  $B_\varphi = (\bar{x}, \bar{y})$  is

$$m_{pq}(f) = \sum_{(x,y) \in \varphi} (x - \bar{x})^p (y - \bar{y})^q f(x, y).$$

Moments provide statistical information about the shape:

- order 0, the surface of the shape:  $A = m_{00}$
- order 1, the centroid of the shape:  $\bar{x} = \frac{m_{10}}{m_{00}}$  and  $\bar{y} = \frac{m_{01}}{m_{00}}$
- and so on

Three families of moments, which are invariant with respect to translation, rotation and homothety are used here to characterize the shapes of nuclei:

- *Hu* moments<sup>20</sup> are seven moments resulting from products and quotients of normalized centered moments of order 3.
- *Legendre* moments<sup>43</sup> are orthogonal moments created with Legendre polynomials. They constitute a complete and orthogonal basis on the interval  $[-1, 1]$ .
- *Zernike* moments<sup>52</sup> are orthogonal complex moments based on a polar representation of the shape.

To characterize the shape, we compute the Legendre and Zernike moments of order 4 and the Hu moments for each nucleus. The subset of moments giving the best classification rate is systematically considered for by cross-validation (see Sec. 2). The winner subset contains 17 moments, and achieves 81% of prediction with neural network (with  $\nu = 3$ ) and cross-validation. Moments of orders 3 and 5 yield results of lower quality.

### 3.2.2. Projection histograms

The *projection histogram* technique<sup>27,45,55</sup> (or *Integral Projection Function*) is frequently used in character recognition. It provides information about the thickness of the shape in various directions. Each histogram is constructed by counting the numbers of pixels in a given direction  $\delta$  so as to get the so called marginal distribution:  $HP_\delta(f) = \sum_\varphi f_\delta(\mathbf{x})$ . Four directions of projections are available in  $\mathbb{Z}^2$ : the horizontal, the vertical and the two diagonals. The method is insensitive to *translation*. On the other hand, it is sensitive to both rotation (it can be fixed by a preliminary rotation according to the major axis), and scale transformation (it is possible to give shapes equal scales since there is a constant ratio between projections).

Subsequently, each nucleus is rotated such that its major axis coincides with the horizontal axis and scaled to be bounded by a box of dimensions  $N \times N$ . Then the horizontal and vertical projection histograms provide  $2N$  characteristics. We have tested this technique for various values of  $N$  for each classification method (by cross-validation). The best result (83%) is obtained by the logistic regression and cross-validation, for a number of characteristics  $N$  equal to 32 (dimensions  $16 \times 16$ ).

### 3.3. Shape indexes

Shape indexes<sup>24</sup> were presented for the first time by Santalo<sup>38</sup> in a book about the mathematical properties of convex shapes. Their definition and properties can be found in Ref. 44.

**Definition 1.** A “shape index” is any parameter, coefficient or combination of coefficients providing numeric information about shape, and satisfying the following

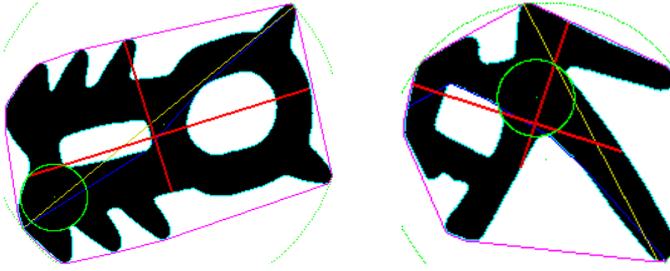


Fig. 2. Examples of measures: surface (black), perimeter (cyan), main and secondary axis (red), convex hull (purple), geodesic diameter (blue) and Euclidian diameter (yellow), smallest (respectively largest) circumscribed (respectively inscribed) ball (green) (color online).

properties:

- (1) No dimension.
- (2) Invariant under scale, rotation and homothety transformations.

**Definition 2.** Each value or set of values “measured” on the shapes is called a “measure”.

The shape indexes are often functions of several measures (see Fig. 2). A complete list of the measures and indexes used in our work is presented in the appendix of this paper. Some well-known measures like the Feret diameter or the radius of maximum (respectively minimum) curvature are not used here because they require too much computation time (high complexity).

The main advantage of shape indexes is their high flexibility. Indeed, it is easy to build new indexes according to the problem to be solved. These specific indexes have a high capacity to describe the aspects of interest of the shape under consideration and are expected to give better classification.

### 3.3.1. Four new shape indexes

### 3.3.2. Indexes for ellipse characterization

Healthy nuclei have a regular and convex shape, close to an ellipse (see Figs. 1(a) and 3(a)). The shape results from the pressure of the cover slip on the microscope slide during observation. It is consequently judicious to build indexes that characterize the elliptic nature of the nuclei.

The area of an ellipse with  $a$  the semi-major axis and  $b$  the semi-minor axis is:  $A = \pi ab$  [see Fig. 3(a)]. An ellipse as the following properties:

- The major axis coincides with the main axis MA (of length  $L_{MA}$ ) and the diameter  $D$ .
- The semi-major axis is equal to the longest radius  $R_{max}$ .
- The minor axis coincides with the secondary axis SA.
- The semi-minor axis is equal to the shortest radius  $R_{min}$  and the thickness from diameter  $E_D$ .

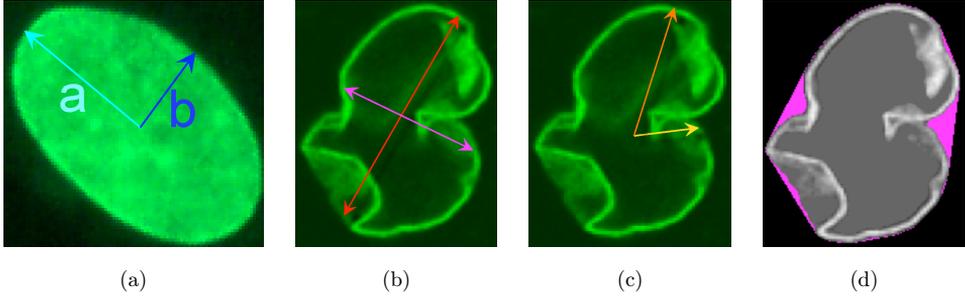


Fig. 3. (a) Illustration of the semi-axes on a nucleus with a normal shape. (b) Main axis (red) and secondary axis (purple). (c) Longest radius (orange) and shortest radius (yellow). (d) Illustration of the calculation of the measure  $N_{Cce}$ , it computes the number of gap connected components (number of purple components) (color online).

We can deduce the following equalities:  $a = \frac{1}{2}L_{MA} = \frac{1}{2}D = R_{\max}$  and  $b = \frac{1}{2}L_{SA} = E_D = R_{\min}$  (see Fig. 3).

Three new shape indexes can be deduced from these equalities:

### Radius

$$\Psi_{\text{Ellipse}R} = \pi \frac{R_{\min} R_{\max}}{A} \in [0, 1]$$

### Main axis

$$\Psi_{\text{Ellipse}MA} = \frac{\pi}{4} \frac{L_{MA} L_{SA}}{A} \in [0, 1]$$

### Diameter

$$\Psi_{\text{Ellipse}D} = \frac{\pi}{2} \frac{E_D D}{A} \in [0, 1].$$

The indexes concerning the radius or the diameter rely on the position of the rim. The indexes constructed from the main axis deal with the surface of the shape. As far as the shape under consideration is a true ellipse, denominators and numerators are equal, so their values are 1. The suggested ranges of variation correspond to convex shapes varying from segment to disk.

#### 3.3.3. Characterization index for convexity

In the previous section, we have constructed three new shape indexes in order to characterize departure from the “normal” ellipsoid shape. A complementary point of view may concern the characterization of shape abnormality.

*Puffy* nuclei contain many concave border areas of various sizes. To evaluate these concave areas, it is possible to calculate the number of connected components  $N_{cc}$  that remain when the shape is subtracted from its convex hull  $N_{cc} = \text{card}(C_H(\varphi) \setminus \varphi)$  [see Fig. 3(d)]. In the following, those connected components are called “gap

components”. A normalized version of the measure is used to elaborate the corresponding index:

$$\Psi_{N_{cc}} = \frac{1}{1 + N_{cc}} \in ]0, 1].$$

The index is equal to 1 if no gap component is found and tends to 0 as the number of gap components in the shape increases.

In practice, components whose area equals one pixel result from resolution errors and cannot be considered as gap components. Even small gap components (i.e. a few pixels) may not be significant, at least with respect to the classification of nuclei. In fact, the sizes as well as the number of gap components must be taken into account when diagnosing nuclei. Thus  $N_{cc}$  requires a calibration to decide whether a gap component should be counted or not. A systematic analysis of the percentage of correct classifications (as observed by index  $\Psi_{N_{cc}}$ ) versus the size and the number of gap components was consequently conducted (see Fig. 4): for a given size  $s$  and a given number  $n$ , we considered that a nucleus has an abnormal shape if it contains at least  $n$  gap components with an area greater than or equal to  $s$ .

A systematic analysis indicates that 95.4% of nuclei with *puffy* shape have at least a gap component of 32 pixels or two gap components of 12 pixels. Note that the flat zone in the resulting surface corresponds to nondiscriminant sizes and/or numbers, which result in the classification of all nuclei in the “normal shape” class (70% of the nuclei considered for this study have a normal shape). Both thresholds are given equivalent weights in  $\Psi_{N_{cc}}$ , so that their combination provides a percentage of

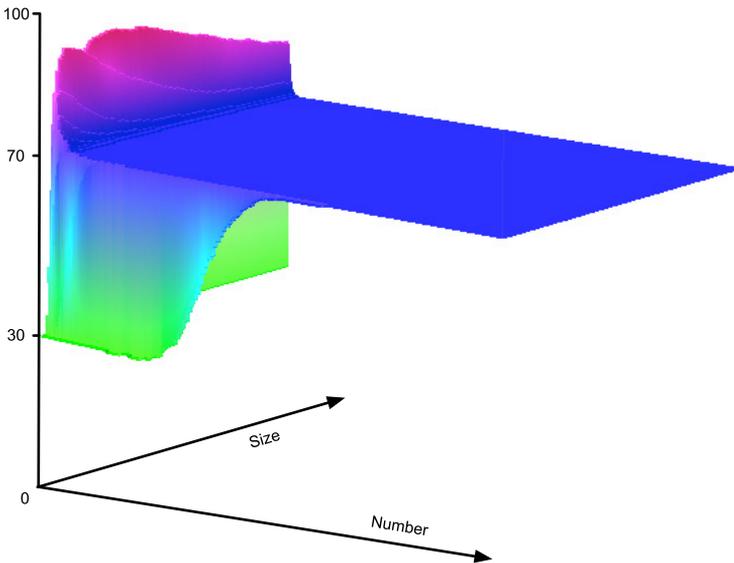


Fig. 4. Surface representing the percentage of classification of nuclei according to the number and the size of gap components.

prediction upper than 90% for the subproblem of shape characterization with this single index.

### 3.4. Classification of cell nuclei shapes

A total of 13 indexes selected from the scientific literature<sup>44</sup> are considered in addition to the new four shape indexes specifically designed for the characterization of the shapes of nuclei. For each classifier, we perform a systematic/exhaustive study (with 10-fold cross validation) in order to find the best subset of indexes (see Fig. 5).

LR and neural networks produce comparable results. For both models, the highest classification rate (95.4%) is obtained with 10 indexes (including the four indexes proposed): the learning and validation errors converge for a number of indexes equal to 10 (results not shown). The probability to get such a result by chance is lower than  $10^{-4}$  and the 95% confidence interval is [95.2, 95.6].<sup>15</sup> The LR is selected over the neural network since it is preferable to use the simplest model with the lowest complexity to get the highest robustness. Without using the four indexes that we elaborated, the best subset is composed of 10 out of the 13 remaining indexes. It provides 93.6% of prediction with [93.3, 93.9] as 95% confidence interval. The result is 2% lower than the one obtained by the previous model composed of our indexes ( $\Psi_{N_{cc}}$  and  $\Psi_{\text{Ellipse}^*}$ ). This result demonstrates the efficiency of the new indexes, since the confidence intervals do not overlap.

The analysis of the distribution of probability given by the classifier (see Fig. 6), reveals that:

- A clear separation of nuclei into 2 groups: nuclei concentrated at both extremities in the histogram.
- A quasi-absence of severe errors: false (errors of classification) with extreme probabilities (lower than 0.2 or upper than 0.8).

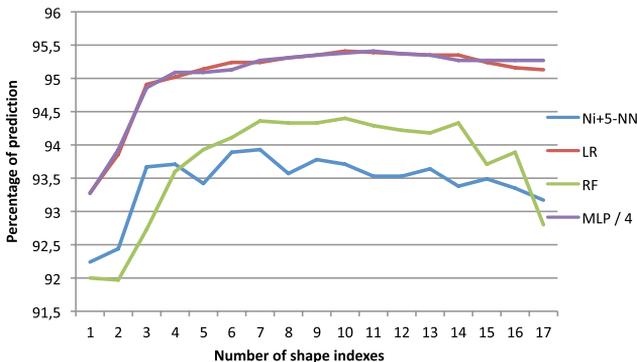


Fig. 5. Comparison of the efficiency of different classification methods: nearest neighbors (NN with  $N_i + 5$ ), logistic regression (LR), RF and MLP with  $\nu = 4$  as a function of the number of indexes (color online).

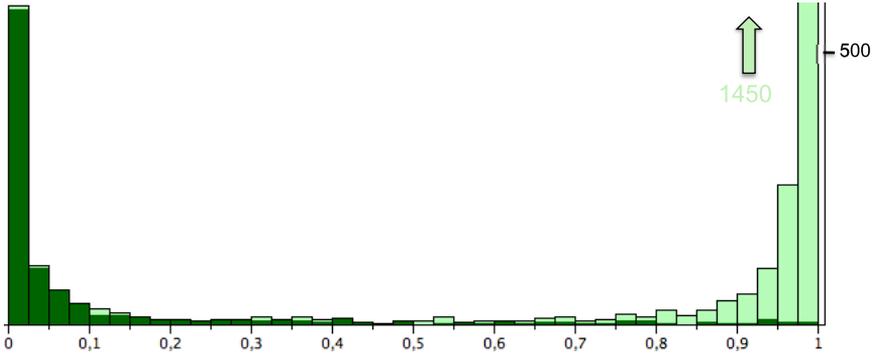


Fig. 6. Probability distribution of nuclei given by the submodel of shape classification. The  $x$ -axis indicates the probability for a nucleus to belong to the group with normal shape. In green (respectively dark green) nuclei with normal shape (respectively puffy) (color online).

- The presence of few ambiguous cases: nuclei with probabilities inside the interval  $[0.3, 0.7]$  (near of the decision value 0.5).

These three characteristics highlight the low level of ambiguity in the classification and confirm the efficiency of the submodel.

## 4. Characterization and Classification of the Texture

### 4.1. Previous work

#### 4.1.1. Shape indexes for texture characterization

Shape indexes have shown their efficiency in the submodel for shape classification. However, Chen *et al.* have developed an approach that allows us to use shape indexes for texture characterization.<sup>7</sup> They exploited the idea that an image is the sum of all binary thresholds:

$$f(\mathbf{x}) = \sum_{\alpha=1}^N f_b(\mathbf{x}, \alpha) \quad \text{with} \quad f_b(\mathbf{x}, \alpha) = \begin{cases} 1 & \text{if } f(\mathbf{x}) \geq \alpha \\ 0 & \text{else} \end{cases}$$

where  $N$  is the number of gray levels of the image, and  $E_\alpha$  is the set of connected components from the thresholded image  $f_b(\mathbf{x}, \alpha)$  and  $\alpha \in [1, N]$  the threshold value. For each value of  $\alpha$  and for a shape index  $\chi$ , we can compute:

$$\chi_\alpha = \frac{\sum_{\varphi \in E_\alpha} [A(\varphi) * \chi(\varphi)]}{\sum_{\varphi \in E_\alpha} A(\varphi)}.$$

The four values *max*, *average*, *weighted average* and *sample standard deviation*, which are calculated for all values of  $\alpha$ , constitute the texture characteristics for the index  $\chi$ .

We used this technique with the indexes composing the best subset of shape indexes previously found (see Sec. 3.4). However, each index produces four values so that overfitting may occur. Moreover, any subset of indexes is not efficient to solve the subproblem of texture classification. The best result (composed of 28 indexes) provides only 85% of prediction.

4.1.2. Gray level cooccurrence matrix and Haralick features

The gray level cooccurrence matrix (COM or dependency spatial matrix) is one of the most well-known, oldest and efficient methods of statistical texture representation. It realizes a second-order statistical analysis, by the study of spatial relationships of pairs of pixels.<sup>16,17</sup>

The COM enumerates existing relationships between the gray levels of pixels in the texture for a given displacement (vector of translation)  $\vec{d}$ . The result is a square matrix of size  $N \times N$ , where  $N$  is the number of gray level of the texture. For a given displacement  $\vec{d} = (d_x, d_y)$ , a matrix element  $M_d(g_x, g_y)$  is the number of pixels that have a gray level  $g_y$  located at  $\vec{d}$  from a pixel of gray level  $g_x$  (see Fig. 7). This can be written as follows:

$$M_d(g_x, g_y) = \text{card}\{((x, y), (x + d_x, y + d_y)) / f(x, y) = g_x, f(x + d_x, y + d_y) = g_y\}.$$

A COM relationships between pixels, according to two aspects: local (gray level) and spatial (displacement). However, when all gray levels and several displacements are considered, it generates a significant amount of information. For a given gray level  $N$ , we compute four matrices for the displacements  $\vec{d}_1 = (1, 0)$ ,  $\vec{d}_2 = (1, 1)$ ,  $\vec{d}_3 = (0, 1)$  and  $\vec{d}_4 = (-1, 1)$ . Then we take the average of the resulting matrices,<sup>48</sup> which allows us to merge information and abstract directions. However, it is always necessary to compute matrices for various displacements (distance  $\varepsilon \vec{d}_i$ ,  $\varepsilon \in \mathbb{N}^*$ ) and gray level reductions, which increases the computing time and quantity of information.

With this reduced matrix, we compute *Haralick's features (second-order texture indexes)*.<sup>16,17</sup> We implement a systematic search in order to find the most efficient classification method and the best subset of texture indexes. Moreover, we perform the computations for 16, 32 and 64 gray levels, and for various distances ( $\varepsilon = 1, \dots, 5$ ). As a consequence of the low number of nuclei in the working set (see Sec. 2), validation is accomplished with the Loo protocol.

Texture		<table style="border-collapse: collapse; width: 100%; text-align: center;"> <tr> <td style="border: 1px solid black; padding: 5px;"><math>N</math></td> <td style="border: 1px solid black; padding: 5px;"><math>\vec{d} = (0, 1)</math></td> <td style="border: 1px solid black; padding: 5px;"><math>\vec{d} = (1, 1)</math></td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;"></td> <td style="border: 1px solid black; padding: 5px;">0   1   2</td> <td style="border: 1px solid black; padding: 5px;">0   1   2</td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">1</td> <td style="border: 1px solid black; padding: 5px;">1   0   0</td> <td style="border: 1px solid black; padding: 5px;">0   1   1</td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">1</td> <td style="border: 1px solid black; padding: 5px;">1   0   0</td> <td style="border: 1px solid black; padding: 5px;">1   1   0</td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">0</td> <td style="border: 1px solid black; padding: 5px;">0   2   2</td> <td style="border: 1px solid black; padding: 5px;">1   0   1</td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">0</td> <td style="border: 1px solid black; padding: 5px;">0   2   2</td> <td style="border: 1px solid black; padding: 5px;">1   0   1</td> </tr> </table>	$N$	$\vec{d} = (0, 1)$	$\vec{d} = (1, 1)$		0   1   2	0   1   2	1	1   0   0	0   1   1	1	1   0   0	1   1   0	0	0   2   2	1   0   1	0	0   2   2	1   0   1
$N$	$\vec{d} = (0, 1)$	$\vec{d} = (1, 1)$																		
	0   1   2	0   1   2																		
1	1   0   0	0   1   1																		
1	1   0   0	1   1   0																		
0	0   2   2	1   0   1																		
0	0   2   2	1   0   1																		
	⇒	<table style="border-collapse: collapse; width: 100%; text-align: center;"> <tr> <td style="border: 1px solid black; padding: 5px;"><math>N</math></td> <td style="border: 1px solid black; padding: 5px;"><math>\vec{d} = (0, 1)</math></td> <td style="border: 1px solid black; padding: 5px;"><math>\vec{d} = (1, 1)</math></td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">0</td> <td style="border: 1px solid black; padding: 5px;">4   0   2</td> <td style="border: 1px solid black; padding: 5px;">2   1   1</td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">1</td> <td style="border: 1px solid black; padding: 5px;">2   2   0</td> <td style="border: 1px solid black; padding: 5px;">1   1   0</td> </tr> <tr> <td style="border: 1px solid black; padding: 5px;">2</td> <td style="border: 1px solid black; padding: 5px;">0   0   2</td> <td style="border: 1px solid black; padding: 5px;">1   0   1</td> </tr> </table>	$N$	$\vec{d} = (0, 1)$	$\vec{d} = (1, 1)$	0	4   0   2	2   1   1	1	2   2   0	1   1   0	2	0   0   2	1   0   1						
$N$	$\vec{d} = (0, 1)$	$\vec{d} = (1, 1)$																		
0	4   0   2	2   1   1																		
1	2   2   0	1   1   0																		
2	0   0   2	1   0   1																		

Fig. 7. Two cooccurrence matrices for a  $4 \times 4$  image with 3 gray levels (left part of the figure), for displacements  $\vec{d} = (1, 0)$  and  $\vec{d} = (1, 1)$ .

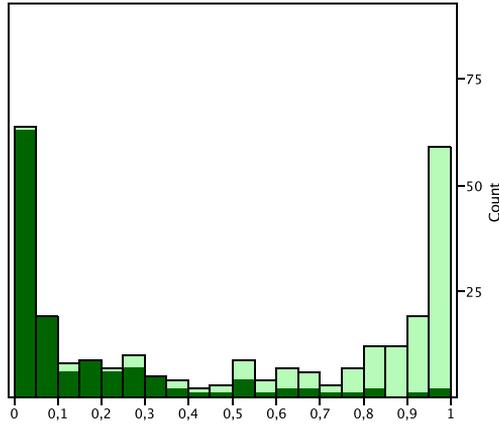


Fig. 8. The probability distribution as given by the model. In light green (respectively dark green), the nuclei with a homogeneous (respectively nonhomogeneous) texture. The nearer the probability to 1 (on the  $x$ -axis), the more homogeneous the texture (color online).

The best classification submodel is provided by LR (neural networks gets comparable results). It is obtained with a matrix computed from 32 gray levels, for a distance of 1 and summarized by 8 features: variance, correlation, average of sums, entropy of sums, entropy, variance of differences, homogeneity and dissimilarity. The submodel obtains 89.8% of prediction with [86.9, 92.7] as 95% confidence interval and a probability lower than  $10^{-4}$ . However, the size of the confidence interval shows that the submodel is sensitive to data: indeed excluded data alter performances. In spite of the high concentration rates at both extremities (as can be seen in Fig. 8), there are 40 ambiguous cases and 8 severe errors (errors with an extreme probability of classification).

#### 4.1.3. Gray level run length matrix

Based on higher-order statistical textural feature, the Gray Level Run Length Matrix ( $\mathcal{RL}$ ) captures the coarseness of the texture.<sup>9,16</sup> It consists in counting the number of pixel segments with the same intensity (flat zone) in a given direction, then representing the results in a matrix. A direction ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  or  $135^\circ$ ) and a number of gray levels are chosen beforehand. The matrix element  $\mathcal{RL}_{\theta,f}(l_n, g_n)$  is equal to the number of segments of length  $l_n$  and gray level  $g_n$  of direction  $\theta$ . This implies that the number of columns in the matrix is dynamic, as it is determined by the length of the longest segment. By design, the calculation is symmetric and therefore, it is unnecessary to consider the four complementary directions ( $180^\circ$ ,  $225^\circ$ ,  $270^\circ$  or  $315^\circ$ ). Figure 9 shows an example of the calculation of a run length matrix.

Once the matrix has been obtained, 11 indexes are calculated<sup>42,50</sup> (moments of order  $-2$  to  $2$ ) to determine the vector that characterizes the texture. To establish our model, the matrix is computed for a given gray level and for four directions. Then, for each index, the average value of the four directions ( $\mathcal{RL}_{0,f}$ ,  $\mathcal{RL}_{45,f}$ ,  $\mathcal{RL}_{90,f}$

Texture			
1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

 $\implies$ 

$g_n$	Run length $l_n$			
	1	2	3	4
1	4	0	0	0
2	1	0	1	0
3	3	0	0	0
4	3	1	0	0

Fig. 9. Example of the computation of a run length matrix for a  $4 \times 4$  texture in the  $0^\circ$  direction for 4 gray levels.

and  $\mathcal{RL}_{135,f}$ ) is considered to avoid dependency on direction and to get a global view of the texture information. A systematic study shows that the best model is obtained for a set of 7 indexes and 32 gray levels. The classification success rate is 84.81% by LR (and neural networks), which is inferior to the rate obtained with the COM summarized by the Haralick’s features (90%).

**4.2. New method: Gray level size zone matrix**

In our hands, homogeneous texture are composed of large flat zones of close gray level, without (or really few) flat zones or segments in any given direction. In light of this, we propose a new approach, which keeps track in a matrix of the size of each flat zone. The resulting matrix is computed according to the run length matrix principle: the value of the matrix element  $\mathcal{GS}_f(s_n, g_n)$  is equal to the number of flat zones of size  $s_n$  and of gray level  $g_n$ . Figure 10 shows an example of the computation of such a matrix, called the *size zone matrix* (SZM).

The resulting matrix has a fixed number of rows (equal to the number of gray levels) and a dynamic number of columns (determined by the size of the largest flat zone). The more homogeneous the texture, the *wider* and *flatter* the matrix. It must be pointed out that this matrix does not require computations in several directions, which are replaced by the labeling of connected components. However, it is still necessary to specify the number of gray levels, but the calculations are subsequently robust to noise. The 11 same indexes are computed the same way as the  $\mathcal{RL}$  matrix with 32 gray levels. The classification rate is 91.11% by LR (and neural networks) with [89.1, 93.1] as the 95% confidence interval of and a probability lower than  $10^{-4}$ . It is the best result obtained by the approaches used so far.

Texture			
1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

 $\implies$ 

Gray Level	Size Zone $s_n$			
$g_n$	1	2	3	4
1	2	1	0	0
2	1	0	1	0
3	0	0	1	0
4	2	0	1	0

Fig. 10. Example of the computation of a in 8-connexity for a  $4 \times 4$  texture in 4 gray levels.

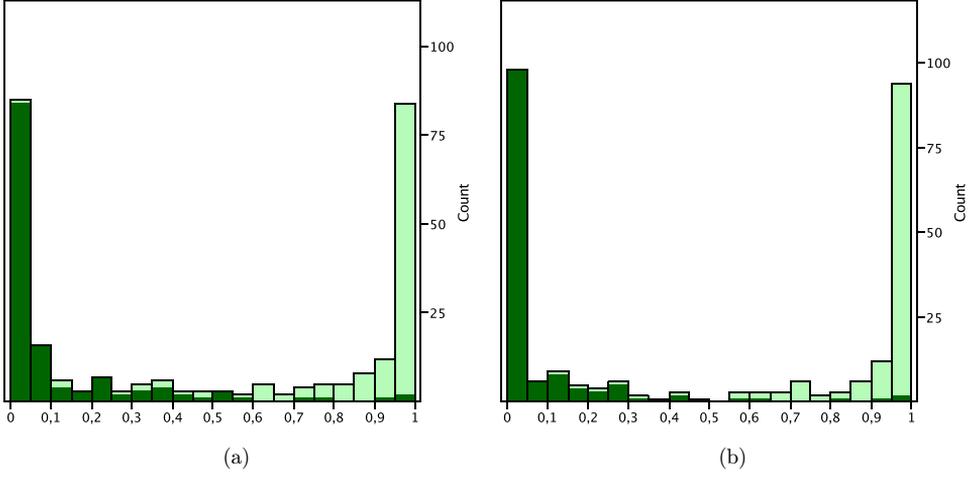


Fig. 11. Probability distribution as given by the models: (a) simple size zone matrix, (b) size zone matrix with the new indexes. In light green (respectively dark green), the nuclei with homogeneous (respectively nonhomogeneous) texture. The nearer the probability to 1 (respectively 0), the more homogeneous (respectively nonhomogeneous) the nuclei's texture.

This improvement can be clearly seen by comparing Figs. 8 and 11(a). A better distribution of the nuclei is visible at both extremities and a reduction of 16 ambiguous cases (only 29 ambiguous cases remain). However, six severe errors remain. These results demonstrate the efficiency of the classification.

However, when examining the data, the indexes and the false positive results, it becomes clear that a specific texture case was not correctly characterized and was the cause of the remaining errors: nuclei with large flat zones, but with high variations in the intensity between flat zones [see Fig. 1(c)], making them inhomogeneously textured nuclei.

To characterize these nuclei types, two new indexes, which are the weighted variances of gray level or area size, are considered:

$$\Psi_N(f) = \sqrt{\frac{1}{N * S} \sum_{g_n=1}^N \sum_{s_n=1}^S (g_n * \mathcal{GS}_f(g_n, s_n) - \mu_N)^2}$$

$$\mu_N = \frac{1}{N * S} \sum_{g_n=1}^N \sum_{s_n=1}^S g_n * \mathcal{GS}_f(g_n, s_n),$$

$$\Psi_S(f) = \sqrt{\frac{1}{N * S} \sum_{g_n=1}^N \sum_{s_n=1}^S (s_n * \mathcal{GS}_f(g_n, s_n) - \mu_S)^2}$$

$$\mu_S = \frac{1}{N * S} \sum_{g_n=1}^N \sum_{s_n=1}^S s_n * \mathcal{GS}_f(g_n, s_n),$$

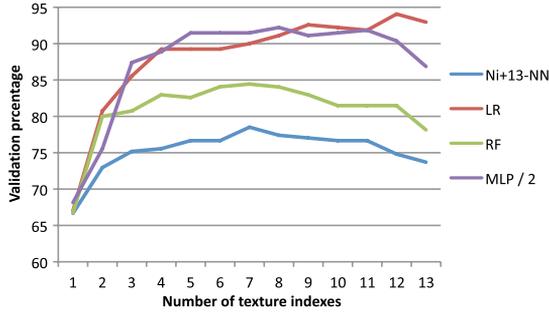


Fig. 12. Comparison of the performance of different classification methods with respect to the number of indexes used: nearest neighbor (NN with  $N_i + 13$ ), logistic regression (LR), random forests (FA) and MLP with  $\nu = 2$ .

where  $N$  and  $S$  are the dimensions of the matrix. The more the texture is made of large flat zones with high intensity variations between them, the higher the value of the index  $\Psi_N$ . In the case of a more homogeneous texture, the value of this index will be low. It is also true for the index  $\Psi_S$ , when dealing with flat zones sizes.

In this way, two new texture indexes are added to the 11 previous ones, making a total of 13 indexes. A systematic study was once again undertaken. This study indicates that the best subset is composed of 12 indexes (only the LRHGLE index discarded) on images reduced to 32 gray levels and classified by LR (see Fig. 12) or neural networks (with  $\nu = 2$ ). Thanks to this new set of indexes, the classification rate of 94.07% has been reached, with a confidence interval of [92.1, 95.9] and a probability lower than  $10^{-4}$ . The use of our indexes improves the submodel’s prediction rate by 3%.

However, there is an intersection between this confidence interval and the previous (without our indexes  $\Psi_N$  and  $\Psi_S$ ). In fact, the probability of the previous submodel to give comparable results to the new submodel is lower than 0.087 (determined by a study of the rank). Moreover, an analysis of variance (Wilcoxon’s test<sup>49</sup>), proves that the probability of these distributions arising from the same law is lower than  $10^{-4}$ . Then our indexes are relevant, necessary and efficient in the submodel.

The improvement provided by our indexes can be observed in Fig. 11(b). The high concentration rate of probabilities at both extremities of the histogram shows the efficiency of the classification and the relevance of the choice of indexes. Moreover, we can observe a higher concentration rate at both extremities and a drop-off in the presence of ambiguous cases (only 16 nuclei remain ambiguous).

A new and all-around recently published method (WND-CHARM)<sup>35</sup> offer the occasion to evaluate the models of classification presented above. WND-CHARM extracts 2700 characteristics of shape (10%) and texture (90%). Next, the best subset of characteristics is selected together with a weighting of the selected characteristics. Finally, a classification based on NN is realized.

We have used this method on the subproblem of texture classification, because it uses a large range of texture characteristics that allows comparison of the efficiency of our approach. The best configuration of this method (using 10% of the most pertinent characteristics) provides a classification rate of 90%. It is a result similar to the one obtained with Haralick's features, which are used in the method.

This last result demonstrates the relevance and the efficiency of our technique (SZM) in the subproblem of texture classification.

## 5. Conclusion

In this paper, the problem of cell nuclei classification by the analysis of their shapes and textures has been addressed and separated into two steps: the first step is based on the study of the shapes, with the design of four new dedicated shape indexes, particularly pertinent and efficient according to the gain of performances obtained. Although these indexes were especially designed for this problem, they can be used to characterize all types of cells, and more generally ellipsoidal and convex shapes. In the second step, a novel statistical method of texture characterization has been presented: it consists in the computation of a representative matrix for the texture, whose efficiency has been compared to another techniques, and more particularly to one that uses a very large range of texture characteristics. The gap of performance was significant.

On both steps of shape and texture classification, the submodels built were systematically validated by cross validation, the confidence interval and the probability were computed. All these information have proven the reliability of submodels.

To build the final classification model, we use the LR (a neural network provides comparable results) in order to classify nuclei according to their probabilities of classification (given by the shape and texture submodels). We obtain a classification percentage of 87.8%, which is lower than the classification percentage of both submodels. This is due to the fact that the majority of nuclei with nonhomogeneous texture (approximately 100 on 135) have a *puffy* shape. The texture classification is thus highly correlated with shape classification. Moreover, only 94% of the nuclei can be diagnosed by shape and/or texture. There remain 6% of the nuclei that can be diagnosed only with rare information, which occurs occasionally. Furthermore, we have used the WND-CHARM algorithm on the final problem. This method was able to classify only 71% of the nuclei, whereas we achieved 88% out of the 94% possible.

In the paper, we have described the different steps necessary to construct an automatic classification model of cell nuclei of patients affected by the Progeria disease. It is now possible to classify cell nuclei automatically, reliably and quickly. The aims of this work were automation, stability and time saving (in comparison with time necessary to a manual classification by experts and its ratio of repeatability), and they have been achieved. Moreover, our work demonstrates the feasibility of classifying cell nuclei by the study of lamins A/C repartition.

In addition, abnormalities in nucleus shape or texture appeared to be a frequent morphological parameter in several human diseases, one of them being cancer. Therefore, the present work will provide a new automatic tool to analyze a large number of nuclei from diseased cells after the detection of different nuclear proteins, being either nuclear skeletal proteins such as lamins A, B and C or more “soluble” proteins involved in several aspects of nuclear metabolism (DNA replication or repair, RNA transcription and splicing).

**List of measures and shape indexes used in this paper**

<b>Extension by diameter</b>	$E_D/D \in [0, 1]$
<b>Extension by radius</b>	$\rho_i/\rho_e \in [0, 1]$
<b>Extension geodesic</b>	$4A/(\pi D_G^2) \in [0, \frac{\pi}{4}]$
<b>Circularity</b>	$R_{\min}/R_{\max} \in [0, 1]$
<b>Convexity by perimeter</b>	$P(C_H)/P \in ]0, 1]$
<b>Convexity by surface</b>	$A/A(C_H) \in ]0, 1]$
<b>Deficit</b>	$\pi(\rho_e - \rho_i)^2/P^2 \in [0, \frac{\pi^2}{16}]$
<b>Iso-perimetric deficit</b>	$4\pi A/P^2 \in [0, 1]$
<b>Gap inscribed disk</b>	$\pi\rho_i^2/A \in [0, 1]$
<b>Spreading of Morton</b>	$4A/(\pi L_{MA}^2) \in [0, 1]$
<b>Irregularity</b>	$(A + \sqrt{\pi} \max_{p \in F} d(p, B))/\sqrt{A}$
<b>Symmetry of Bezicovitch</b>	$\max_{p \in F} A(F \cap \text{Symmetric}_p(F))$
<b>Circularity of variance</b>	$\frac{1}{ \text{Rim}(F) \mu_R^2} \sum_{p \in \text{Rim}(F)} (\ p - B\  - \mu_r)^2$

with :

- $A$  surface.
- $B$  barycenter.
- $C_H$  convex hull.
- $D$  diameter.
- $D_G$  geodesic diameter.
- $E_D$  thickness from diameter.
- $L_{MA}$  length of the main axis and respectively.
- $L_{SA}$  the length of the secondary axis (orthogonal to main axis).
- $P$  perimeter.
- $\rho_e$  radius of the smallest circumscribed sphere.
- $\rho_i$  radius of the largest inscribed sphere.
- $R_{\min}$  shortest radius.
- $R_{\max}$  longest radius.
- $\mu_r$  average radius.

## References

1. Y. Amit and D. Geman, Shape quantization and recognition with randomized trees, *Neural Comput.* **9** (1997) 1545–1588.
2. J. Berkson, Application of the logistic function to bio-assay, *J. Am. Stat. Assoc.* **39** (1944) 357–365.
3. M. Blumensteina, X. Y. Liua and B. Verma, An investigation of the modified direction feature for cursive character recognition, *Pattern Recognit.* **40** (2007) 376–388.
4. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification And Regression Trees* (CRC Press, 1984).
5. L. Breiman, Bagging predictors, *Mach. Learn.* **24** (1996) 123–140.
6. L. Breiman, Random forests, *Mach. Learn.* **45** (2001) 5–32.
7. Y. Q. Chen, M. S. Dixon and D. W. Thomas, Statistical geometrical features for texture classification, *Pattern Recognit.* **28** (1995) 537–552.
8. G. Y. Chen and Balázs Kegl, Invariant pattern recognition using dual tree complex wavelets and fourier features, *Pattern Recognit.* **42**(9) (2009) 2013–2019.
9. A. Chu, C. M. Sehgal and J. F. Greenleaf, Use of gray value distribution of run lengths for texture analysis, *IEEE Trans. Pattern Recognit. Lett.* **11** (1990) 415–419.
10. T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Comput.* **10** (1998) 1895–1923.
11. M. Eriksson, T. W. Brown, L. Gordon, M. W. Glynn, J. Singer, L. Scott, M. R. Erdos, C. M. Robbins, T. Y. Moses, P. Berglund, A. Dutra, E. Pak, S. Durkin, A. B. Csoka, M. Boehnke, T. W. Glover and F. S. Collins, Recurrent de novo point mutations in lamin a cause Hutchinson-Gilford progeria syndrome, *Nature* **423** (2003) 6937.
12. E. Fix and J. Hodges, Discriminatory analysis: Nonparametric discrimination: Consistency properties, *USAF Sch. Aviat. Med.* **4** (1951) 261–279.
13. J. Flusser and T. Suk, Affine moment invariants: A new tool for character recognition, *IEEE Trans. Pattern Recognit. Lett.* **15** (1994) 433–436.
14. J. Flusser and T. Suk, Rotation moment invariants for recognition of symmetric objects, *IEEE Trans. Image Process.* **12** (2006) 3784–3790.
15. B. K. Gosh, A comparison of some approximate confidence intervals for the binomial parameter, *J. Am. Stat. Assoc.* **74** (1979) 894–900.
16. R. M. Haralick, K. Shanmugam and I. Dinstein, Textural features for image classification, *IEEE Trans. Syst. Man Cybern.* **3** (1973) 610–621.
17. R. M. Haralick, Statistical and structural approaches to texture, *Proc. IEEE* **67** (1979) 786–804.
18. J. A. Hartigan and M. A. Wong, A K-Means clustering algorithm, *Appl. Stat.* **28** (1979) 100–108.
19. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression* (John Wiley & Sons, Toronto, 1989).
20. M. K. Hu, Visual pattern recognition by moment invariant, *IEEE Trans. Inf. Theory* **8** (1962) 179–187.
21. N. Japkowicz, Learning from imbalanced data sets: A comparison of various strategies, *AAAI Workshop on Learning from Imbalanced Data Sets*, 2000, pp. 10–15.
22. N. Japkowicz, The Class imbalanced problem: Significance and strategies, *Int. Conf. Artificial Intelligence (ICAI)*, 2000, pp. 111–117.
23. K. Kpalma and J. Ronsin, Multiscale contour description for pattern recognition, *IEEE Trans. Pattern Recognit. Lett.* **27** (2006) 1545–1559.
24. N. Lassouaoui, L. Hamami and N. Nouali, Morphological description of cervical cell images for the pathological recognition, *World Acad. Sci. Eng. Technol.* **5** (2007) 49–52.

25. X. Y. Liu, J. Wu and Z. H. Zhou, Exploratory under-sampling for class-imbalance learning, *IEEE Int. Conf. Data Mining*, Washington, DC, USA, 2006, pp. 965–969.
26. A. Liu, J. Ghosh and C. Martin, Generative oversampling for mining imbalanced datasets, *IEEE Int. Conf. Data Mining*, eds. R. Stahlbock, S. F. Crone and S. Lessmann, Washington, DC, USA, 2007, pp. 66–72.
27. L. M. Lorigo and V. Govindaraju, Off-line arabic handwriting recognition: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.* **28** (2006) 712–724.
28. S. Marcellin, D. A. Zighed and G. Ritschard, An asymmetric entropy measure for decision trees, *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2006, pp. 1292–1299.
29. J. L. Mari, Approaches dedicated to the modeling of complex shapes — Application to medical data, *J. Med. Inform. Technol.* **7** (2004) 3–15.
30. H. A. Martens and P. Dardenne, Validation and verification of regression in small data sets, *Chemometr. Intell. Lab. Syst.* **44** (1998) 99–121.
31. W. S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5** (1943) 115–133.
32. D. McFadden, Conditional logit analysis of qualitative choice behaviour, in *Frontiers in Econometrics*, ed. P. Zarembka (Academic Press, NY, 1974), pp. 105–142.
33. T. Ojala and M. Pietikäinen and D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, *Int. Conf. Pattern Recognition (ICPR)*, Vol. 1, 1994, pp. 582–585.
34. T. Ojala and M. Pietikäinen and D. Harwood, A comparative study of texture measures with classification based on featured distributions, *Pattern Recognit.* **29**(1) (1996) 51–59.
35. N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley and I. G. Goldberg, Multi-purpose image classification using compound image transforms, *Pattern Recognit. Lett.* **29** (2008) 1684–1693.
36. F. Rosenblatt, The Perceptron: Probabilistic model for information storage and organization in the brain, *Psychol. Rev.* **65** (1958) 386–408.
37. A. De Sandre-Giovannoli, R. Bernard, P. Cau, C. Navarro, J. Amiel, I. Boccaccio, S. Lyonnet, C. L. Stewart, A. Munnich, M. Le Merrer and N. Levy, Lamin a truncation in progeria, *Science* **5628** (2003) 2055.
38. L. A. Santalo, *Integral Geometry and Geometric Probability* (Addison Wesley, 1976).
39. H. Soltanzadeh and M. Rahmati, Recognition of Persian handwritten digits using image profiles of multiple orientations, *IEEE Trans. Pattern Recognit. Lett.* **25** (2004) 1569–1576.
40. M. Stone, Cross-validatory choice and assessment of satatistical predictions, *J. R. Stat. Soc.* **36** (1974) 111–147.
41. G. N. Srinivasan and G. Shobha, Statistical texture analysis, *World Acad. Sci. Eng. Technol.* **36** (2008) 1264–1269.
42. X. Tang, Texture Information in run-length matrices, *IEEE Trans. Image Process.* **7**(11) (1998) 1602–1609.
43. M. R. Teague, Image analysis via the general theory of moments, *J. Opt. Soc. Am.* **70** (1980) 920–930.
44. G. Thibault, C. Devic, J. F. Horn, B. Fertil, J. Sequeira and J. L. Mari, Classification of cell nuclei using shape and texture indexes, *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2008, pp. 25–28.
45. O. D. Trier, A. K. Jain and T. Taxt, Recognition of Persian handwritten digits using image profiles of multiple orientations, *IEEE Trans. Pattern Recognit. Lett.* **29** (1996) 641–662.

46. M. Tuceryan and A. K. Jain, Texture analysis,” in *The Handbook of Pattern Recognition & Computer Vision* (World Scientific Publishing Co., Inc. J. Winker, River Edge, NJ, 1998), pp. 207–248.
  47. M. Unser, Texture classification and segmentation using wavelets frames, *IEEE Trans. Image Process.* **4** (1995) 1549–1560.
  48. G. Van De Wouwer, P. Scheunders and D. Van Dyck, Statistical texture characterization from discrete wavelet representations, *IEEE Trans. Image Process.* **8** (1999) 592–598.
  49. T. H. Wonnacott and R. J. Wonnacott, *Statistique: Economie — Gestion — Sciences — Médecine (avec exercices d’application)*, 4th edn. Economica, 1998.
  50. D. H. Xu, A. S. Kurani, J. D. Furst and D. S. Raicu, Run-Length encoding for volumetric texture, *Int. Conf. Visualization, Imaging and Image Processing (VIIP)*, 2004, pp. 452–458.
  51. C. Zahn and R. Roskies, Fourier descriptors for plane closed curves, *IEEE Trans. Comput.* **21**(3) (1971) 269–281.
  52. F. Zernike, Diffraction theory of the cut procedure and its improved form, the phase contrast method, *Physica* **1** (1934) 689–704.
  53. D. Zhang and G. Lu, A comparative study on shape retrieval using fourier descriptors with different shape signatures, *J. Vis. Commun. Image Represent.* **14** (2003) 41–60.
  54. J. Zhang and T. Tan, Brief review of invariant texture analysis methods, *Pattern Recognit.* **35** (2002) 735–747.
  55. Z. H. Zhou and X. Geng, Projection functions for eye detection, *Pattern Recognit.* **37** (2004) 1049–1056.
- 



**Guillaume Thibault** received his degree in Computer Science, and his Ph.D. in Image Processing and pattern recognition from the University of Aix-Marseille II, France, in 2009. He was formerly a research engineer in the Center of Mathematical Morphology (Department of Mathematics and Systems) at MINES ParisTech, and he is currently a researcher engineer in the Center for Spoken Language Understanding (CSLU) at Oregon Health & Science University.

His research interests are in the areas of image processing, pattern recognition (shape and texture characterization and classification), with applications to biomedical image analysis: cellular biology (cell analysis, mitosis detection, computer aided diagnosis, 2.5D confocal images projection, segmentation) and ophthalmology (main structures segmentation and pathologies detection).



**Bernard Fertil** received his Ph.D. in Physics from Paris XI University (France) in 1975 and in Living Sciences from Paris VI University (France) in 1984. He is currently Research Director at the CNRS Institute, UMR 6168, LSIS Laboratory in Marseille (France). His research interests include data mining, image analysis, modeling, with specific applications to medical decision making, bioinformatics and radiobiology.



**Claire Navarro** obtained in 2007, her Ph.D. in Eukaryote Biology from the Aix-Marseille University, under the supervision of Nicolas Lévy. Then, she obtained a fellowship from University of Milan for a post-doctoral position at “Ospedale Maggiore Policlinico di Milano” in the

Stem cell Laboratory, directed by Yvan Torrente for two years. Since 2010, she has been a post-doctoral researcher in University of Aix-Marseille (AMU), UMR 910, “Génétique Médicale et Génomique Fonctionnelle”. Her research focuses on Progeria, a rare genetic disorder characterized by features of premature aging. She is especially interested by the physiopathology of the disease and involved in new therapeutic approaches developed on a mice model mimicking this devastating disorder.



**Sandrine Pereira** is Ph.D. in the team of Radiobiology since December, 2011. Having supported her doctoral thesis (University of the méditerranée) in September, 2005, she realized 2 post-doctorate fellowships that allowed her to specialize in domains going of

the molecular biology to the OMICS; within the framework of an agreement of collaboration INSERM/IRSN, she is untied in the laboratory of Ecotoxicology of Radionucléides on the site of the CEA CADARACHE. Sandrine Pereira is a researcher who specialized not only in genetics, and neurobiology but also in radiobiology. Her work was rewarded on numerous occasions (Grant for the International Congress of Epilepsy 2004, prize Jean Frézal in 2008; prize APRAT in 2009).



**Pierre Cau**, M.D., Ph.D. is full Professor in Cell Biology at Marseille's Médical Faculty and Head of the Cell Biology Laboratory at La Timone Hospital (Assistance Publique-Hôpitaux de Marseille).



**Nicolas Lévy** is Professor of Human and Molecular Genetics. He is the head of the Medical Genetics department and of the Inserm research center “Medical Genetics and Functional Genomics”. In 2009, he was appointed as the director of the French National Institute for rare

diseases and, in 2012, he created and was appointed as the head of the “rare diseases foundation”. After having identified the *LMNA* mutation causing Progeria, Nicolas Lévy's efforts have been dedicated to gene's identification in related progeroid syndromes and the understanding of pathophysiological mechanisms involved in Progeria. In collaboration with Carlos Lopez-Otin, his laboratory has developed and pre-clinically tested a combination of global prenylation inhibitors in human cells and living mouse model. Based on the demonstrated efficacy of this combination, Nicolas Lévy is the PI of the first phase II trial using amino-bisphosphonates and Statins for Progeria, in progress at La Timone Hospital in Marseille, France. Based on a newly generated mouse model of Progeria, he has developed a gene therapy approach as a potential treatment for Progeria. Recently, his work dedicated to the understanding of mechanisms, and to the development of therapeutic strategies in neuromuscular disorders, led his team to demonstrate the proofs of principle for exon skipping and minigene transfer as potential therapies in rare neuromuscular disorders.



**Jean Sequeira** was born in Marseilles (France) in 1953. He graduated from “Ecole Polytechnique” in 1977 and from “Ecole Nationale Supérieure des Télécommunications” in 1979. Then, he taught Computer Science from 1979 to 1981 in an Ivory Coast engineering school.

From 1981 to 1991, he developed research works in Computer Graphics at the IBM Paris Scientific Center. During this period, he obtained a “Docteur Ingénieur” degree (Ph.D. — 1982) and a “Doctorat d’Etat” degree (Accreditation to research supervision) in 1987. He has been a “Full Professor” at the University of Marseilles since 1991 and he has been an “Exceptional Class Professor” since 2010. He is the manager of the SimGraph research group on “Image Analysis and Simulation” in the LSIS laboratory (Laboratoire des Sciences de l’Information et des Systèmes). He has published more than 120 papers, most of them in journals and international conferences, and he was the Scientific Director of 23 Ph.D. research works. In 2006, he participated in the creation of the “International Society for Digital Earth” (ISDE) and of the “International Journal of Digital Earth” (IJDE — Taylor & Francis). He is in charge of international projects, especially with China, and he is involved in journal editorial boards, conference organization and scientific associations. He is an IEEE Senior Member.



**Jean-Luc Mari** Mari received his Ph.D. in 2002. He has been an Associate Professor since 2003 in the Department of Computer Science at the Faculté des Sciences de Luminy (Université de la Méditerranée) and then at the Faculté des Sciences (Aix-Marseille Université) since 2012. He received his Habilitation à Diriger des Recherches (accreditation to supervise research) in 2011. He is also a member of the Information and System Science Laboratory (LSIS), in the team G-Mod (Geometrical Modeling group). His research interests include geometrical modeling, discrete geometry, feature extraction, model representation, implicit and subdivision surfaces, meshes, multiresolution, skeleton based objects and 3D reconstruction.