

Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification

Guillaume Thibault, Jesús Angulo and Fernand Meyer

Abstract—This paper presents new structural statistical matrices which are gray level Size Zone Matrix (SZM) texture descriptor variants. The SZM is based on the co-occurrences of size/intensity of each flat zone (connected pixels with the same gray level). The first improvement increases the information processed by merging multiple gray levels quantizations and reduces the required parameters number. New improved descriptors were especially designed for supervised cell texture classification. They are illustrated thanks to two different databases built from quantitative cell biology. The second alternative characterizes the DNA organization during the mitosis, according to zone intensities radial distribution. The third variant is a matrix structure generalization for the fibrous textures analysis, by changing the intensity/size pair into the length/orientation pair of each region.

Index Terms—Texture Characterization and Classification, Structural Statistical Matrices, Gray level Size Zone Matrix (SZM), Quantitative Cytology.

I. INTRODUCTION

PARALLEL cells growing in multi-well plates current technologies (or in other supports as cell on chip) and fluorescent labeling of targeted proteins (antibodies immuno-fluorescence, GFP-tagged proteins), are coupled together for automated microscopy image capture and subsequent cell image analysis. All this is essential for new cellular biology mechanisms discovery (i.e., using siRNA), new pharmaceuticals (i.e., potential active molecules mass screening) or for new diagnostic/prognostic tests development, and also toxicology tests (i.e., different concentration compounds assessments). The more cells are acquired, the more accurate analysis is performed. Currently, most of these processes are manual, time consuming and involving results variability according to experts (inter-observer variability).

Cell classification is a pattern recognition classical task [1], [2], but the key point for such a cell classification system is to achieve a high robust throughput system which will be able to automatically analyze thousands of cell images without any manual interaction [3]. Texture characterization and classification are traditionally one of the useful techniques to describe cell features.

Methods presented in this paper were designed to efficiently characterize various significant cells and nuclei

Manuscript received ??, ?; revised ??, 2013.

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

aspects. They allow to describe either the different mitosis phases which are relevant to facilitate the oncology molecule active effect analysis, as the various cells pattern, and are fundamental to search antibodies in the patient serum, or to reveal autoimmune diseases. More precisely, this paper aim is to present several new alternative bivariate statistical texture representations, as extended and completed preliminary works versions presented in [4] and [5]. The first descriptor, discussed in Section IV, is a gray level Size Zone Matrix (SZM) multi-scale extension, which merges various gray levels quantizations and which avoids selecting an “optimal” quantization. As we empirically prove in the paper, this new descriptor improves texture homogeneity characterization. SZM and the new multi-scale extension are particularly efficient to characterize homogeneity. Then two other versions are introduced to characterize other texture types. The second alternative suggested in Section V, takes into account zone intensities radial distribution for DNA characterization. The third variant, studied in Section VI is a matrix structure generalization allowing fibrous textures analysis, by changing the intensity/size pair into length/orientation pair. It is particularly used in order to characterize microtubule network. These new and improved descriptors interests are illustrated all along this paper for texture classification problems arising from quantitative cell biology. Indeed the present paper includes a more rigorous presentation of our descriptors than [4], as well as a more systematic study of their performances for cell classification, including comparative evaluation with other statistical matrix-based texture descriptors. Results obtained using the different databases are presented in Section VII. This paper is also completed with material background. On one hand, the Section II summarizes supervised classification algorithms used in our experiments. On the other hand, the Section III provides a state-of-the-art on previous statistical matrices particularly relevant for cell classification.

Two datasets were used to illustrate our texture descriptors performances. The first cell dataset was provided by the ICPR 2012 *HEp-2 Cells Classification* contest [6]. Cell images were acquired by Indirect ImmunoFluorescence (IFF), using a fluorescence microscope (40-fold magnification), coupled with a 50W mercury vapor lamp and with a digital camera CCD (SLIM system by Das srl), with $6.45\mu\text{m}$ square pixel. It contains 1457 cells divided into 6 classes:

- **Centromere** (388), several discrete speckles distributed throughout the interphase nuclei and characteristically

found in the mitosis condensed nuclear chromatin as a

bar of closely associated speckles.

- **Coarse** (239) or **fine speckled**, granular nuclear staining of the interphase cell nuclei.
- **Cytoplasmic** (128), fine fluorescent fibers running the length of the cell.
- **Homogeneous** (345), diffuse both nuclei interphase and mitotic cells chromatin staining.
- **Nucleolar** (257), large coarse speckled staining within the nucleus, less than six in number per cell.

The second cell dataset is part of the RAMIS project¹. It is composed of thousands 2D z-stacks images, acquired by a AxioImagerZ1 epifluorescent confocal microscope with 63X objective N.A. 1.4 and 12 bits Hamamastu CCD camera, driven by Visilog software. The cell image pre-processing involves the following steps: (i) a z-stacks projection for extended-depth-of-focus 2D images; (ii) an image normalization for an uneven illumination correction [7]; (iii) a microtubule network enhancement (see section VI) and; (iv) an individual cell segmentation, including a nuclei/cytoplasm separation. A selection of 300 individual cells composes this dataset, where each cell was completely annotated by experts (approximately 50 features per cell), including typical annotations on: DNA homogeneity, DNA quantization, microtubule network organization, etc.

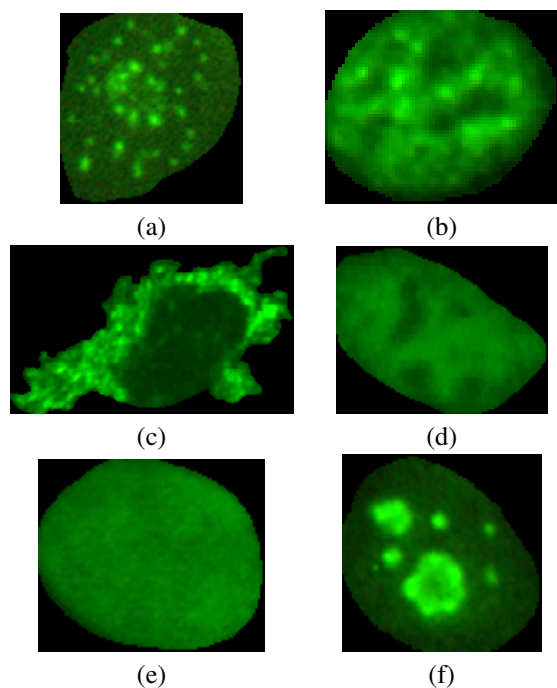


Fig. 1. Examples of typical cells for each category of ICPR'12 cell classification contest: (a) centromeres, (b) coarse speckle, (c) cytoplasmic, (d) fine speckle, (e) homogeneous and (f) nucleolar.

II. CLASSIFICATION

The classification aim is to attribute classes (sometimes just one) to each studied object. In this study we take benefit

¹<http://www.adcis.net/fr/Applications/Projets-De-Recherche-Et-Developpement/RAMIS-Selection-De-Molecules-Innovantes-Inhibant-La-Division-Cellulaire.html>

of biologists knowledge who specified the classes, which are important clues in mitosis phases cell classification and cells pattern. We can therefore take advantage of supervised methods to design the classifiers. Such a classifier is usually built using a learning method, and generalization is achieved thanks to cross-validation. With this objective, datasets are separated into two groups: a training sample and a validation sample. The classifier must have comparable performance levels through training and validation. But prior to the classification phase, it is necessary to construct a characteristics vector describing the data. The vector must be relevant to allow an accurate classification and prediction. The major risk in providing the classifier too many characteristics is *over-fitting*. The greater the characteristic vector dimension, the greater the model flexibility and the better the classification, but the poorer the model's performance for a data set not used during the training. Then each model must systematically be validated and the best classification with the validation sample identified. In this paper, the validation is done following the *K-Fold Cross Validation* [8]–[10] or if necessary with the *Leave-One-Out* protocol [8], [11] (a k -fold validation with k equals to the working set size) because of the working set small size.

In our paper we consider three of the most popular and efficient classification methods, which are formalized through different machine learning paradigms:

- The Logistic Regression (LR) [12]–[14] is a linear model particularly well adapted to classification problems with two classes: $P = P(Y/\vec{x}) = \frac{e^{f(\vec{x})}}{1+e^{f(\vec{x})}}$ with $\vec{x} = (x_1, \dots, x_n)$ being the input data characteristic vector, $f(\vec{x}) = \sum_i \alpha_i x_i$ and $P(Y/\vec{x})$ the conditional probability P of the variable \vec{x} to belong to the class Y . To estimate the model coefficients α_i , the maximum likelihood method is often used, which maximizes the probability of obtaining values observed on the learning sample. It consists in finding parameters optimizing the likelihood function. Logistic regression is preferred to discriminant analysis [15] because of its variables fewer restrictions and its results easier interpretation.
- The Random Forests (RF) [16] is a non-linear classification technique based on the use of *Classification And Regression Trees* (CART) [17]. It is one of the most recent developments in the randomized decision trees aggregation research. It synthesizes approaches developed in [18] and [19].
- The Neural networks [20] (NN) is a non-linear technique where training consists in minimizing the average squared error associated cost using a gradient descent (back-propagation on multilayer perceptrons).

We have access to cells sets labeled by experts, but with imbalanced data sets, specially when there is more than two classes per problem. In order to simplify the characteristic space boundary decision determination, we use a one class classifier technique: the treated class against all the others merged in a single one. Therefore one classifier per class is built, and results are significantly improved. However, in this

case, the imbalanced data set problem is amplified. Thus, we implement an *over-sizing* protocol (or *over-sampling*) [21]: we multiply the minority class in order to obtain two balanced classes. Please note that other techniques are available to deal with imbalance data. It is possible to implement *under-sampling* [22]: random or directed instances suppression in the majority class until the sets are balanced. However it is not recommended for a small data set. Other techniques based on asymmetric entropy measure [23] or the use of an auto-associator neural network [24] could also be used as well.

In order to validate our results, we systematically compute each model confidence interval (CI) and probability. The confidence interval contains 95% of results provided by a model, because it could occur some significant result variance due to the data repartition in learning and training sets. The model probability is the probability to obtain a similar result if predictions are randomly chosen. It is computed by a random class mix of each data in the learning set. These probabilities will be systematically lower than 0.0001, due to the proposed algorithms relevance and efficiency.

III. PREVIOUS WORKS ON STATISTICAL MATRICES

Statistical matrices were extensively used for texture characterization. The most famous one is the gray level Co-Occurrence Matrix (COM), coupled with Haralick's features [25]. The COM represents the texture by second order statistics: distribution of co-occurring values at a given offset. The more offsets used, the more information quantity extracted. This is the main approach drawback, but this method is still a reference in cell texture classification issues [26]–[28].

Another classical technique is the gray level Run Length Matrix (RLM) [29], extensively developed for texture classification [30], [31]. The RLM extracts statistical higher order features: the matrix element $p(i, j|\theta)$ gives the intensity i and length j total runs number (i.e., collinear pixels with same intensity in the same direction θ). This method is particularly efficient for periodic textures and completes the COM informations. Typical features extracted in RLM are moments of order -2 to 2 .

Others statistical matrices were proposed in the state-of-the-art, such as the Gray Level Difference Histogram (GLDH) [32] and the Gray Level Sum Histogram (GLSH) [33], however these methods extract less information and deserved less interest for real applications in quantitative cytology.

IV. MULTIPLE GRAY LEVEL SIZE ZONE MATRIX

Let $f : \begin{cases} E \rightarrow \mathcal{T} \\ \mathbf{x} \mapsto f(\mathbf{x}) \end{cases}$ be a gray-levels image, where $E \subset \mathbf{Z}^2$ is the pixels space support 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 an 8 bits image we have $t_1 = 1$, $N = 256$ and $\Delta t = 1$. Let us also assume that the image f is segmented into its J flat zones $R_j[f]$ (i.e., connected regions of constant value): $E = \cup_{j=1}^J R_j[f]$, $\cap_{j=1}^J R_j[f] = \emptyset$. The size (surface area) of each region is $s(j) = |R_j[f]|$ ($|\cdot|$ is the cardinal). Hence, we consider that each zone $R_j[f]$ has associated a constant gray-level intensity.

A. Reminder on gray level Size Zone Matrix (SZM)

Our starting point is the gray level Size Zone Matrix original notion, based on each flat zone (connected pixels with the same gray level) size/intensity co-occurrences. It was recently introduced in [4], [34] as an alternative to the jointed gray level-run length distribution.

The texture image f SZM, denoted \mathcal{GS}_f , provides a statistical representation by the estimation of a bivariate conditional probability density function of the image distribution values. It is calculated following the pioneering RLM principle: the $\mathcal{GS}_f^N(s_n, g_m)$ matrix value is equal to the size s_n and gray level g_m total zones number, after reduction of f to N gray levels. The resulting matrix has a fixed rows number equals to N (so matrix height), and a dynamic columns number (so matrix width), determined as the largest zone size as well as the quantization size.

The image gray levels (resp. sizes) number can be reduced by a function (generally linear, but also Log, Square Root, etc. according to the problem under investigation) in order to improve the result efficiency and stability. Indeed, two zones of gray level (resp. size) g_m and $g_n = g_m + 1$ (resp. s_n and $s_m = s_n + 1$) are they really different?

Thanks to this design, the more homogeneous the texture (large flat zones with closed gray levels), the wider and flatter the matrix. Upon this statistical and matrix representation, we can calculate all \mathcal{GS}_f second-order moments as compact texture features [30] and two more features which are specific weighted variances [34]. Figure 2 shows an example of such a matrix calculation.

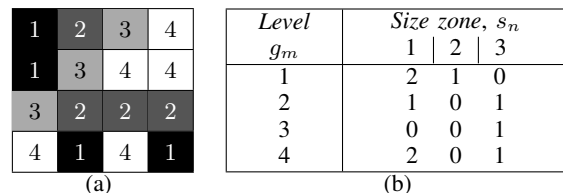


Fig. 2. SZM filling example for a 4×4 image texture dimensions with 4 gray levels and using 8-connectivity.

This matrix does not require a several directions computation, contrary to RLM and COM; however it requires a flat zone labeling which is time consuming. The connectivity type used for labeling modifies the matrix, but our experiments do not show classification performance impact. The RLM and COM are appropriated for periodic textures whereas the SZM is typically adapted to describe non periodic heterogeneous textures. In addition, due to the intrinsic segmentation, texture description in SZM is more regional than the COM point-wise-based representation. However, it was empirically shown [4] (see tables 5 and 7, and proved in the past for other databases) that the gray level quantization degree still has an important impact on the texture classification performance. For a general application it is usually required to test several gray level and width quantizations in order to find the optimal one with respect to a training dataset.

B. Multiple gray levels quantization of SZM

Instead of optimizing the gray levels number N , we propose to construct a multiple scheme with various matrices and then to combine them into a single matrix. The multiple gray level Size Zone Matrix (MSZM) principle, for a 8 bits image, consists in calculating 8 SZM for 8 different gray levels quantizations: $N_k = 2^k$, $k = 1, 2, \dots, 8$, and merge the resulting matrices with a weighted average (which avoids to multiply the matrices number and thus characteristics, and then provides simpler classifiers):

$$\widetilde{\mathcal{G}}\mathcal{S}_f(s_n, g_m) = \sum_{k=1}^8 w_k \mathcal{G}\mathcal{S}_f^{N_k}(s_n, g_m)$$

where $\mathcal{G}\mathcal{S}_f^{N_k}$ is calculated from \mathcal{T} quantized in N_k gray levels. Weights distribution in MSZM is given by a Gaussian function (so $\sum_k w_k = 1$) centered between $N_4 = 16$ and $N_5 = 32$ gray levels: this distribution penalizes gray levels number extreme values because low levels contain limited structural information and high levels are sensitive to noise. By the way, the weights could be automatically learnt or adapted a priori for a specific application. For example, w_k can be equal to $\mathcal{G}\mathcal{S}^{N_k}$ classification rate: for each N_k gray level quantization, $\mathcal{G}\mathcal{S}^{N_k}$ performances are evaluated among the whole dataset and then used as weights. Results are then strongly improved, but it is more time consuming. For N_k values, $\mathcal{G}\mathcal{S}_f^{N_k}$ matrices have different dimensions. Even if we consider that the regions size is quantized with the same intervals (same number of columns), the number of rows is equal to N_k . To solve this drawback, we propose to replicate each of the N_k rows in order to finally obtain 256 rows.

The different gray-levels quantization can also be interpreted as a segmentation into λ -flat zones [35], where the value of λ is associated to the corresponding $\Delta t = 256/N_k$. From a computational viewpoint, the multiple SZM $\widetilde{\mathcal{G}}\mathcal{S}_f$ requires to fill 8 individual SZM, but it is generally more efficient for texture classification (see application Section VII).

V. GRAY LEVEL DISTANCE-TO-BORDER ZONE MATRIX

Texture in natural objects is often non stationary in space; for instance, the texture can radially vary with respect to the object center or border. For instance, we have to characterize the cell nuclei DNA organization (i.e., chromatin texture), see Fig. 7. More precisely, the ‘‘DNA quantity’’ is represented by the pixel intensity: the higher the pixel gray level, the higher the DNA quantity (this is one of the reasons why we previously corrected the uneven illumination). Observing the examples, the DNA distribution is not stationary and, for some classes, it is usually further to the nuclei border (according to the skeleton, see Fig. 7 d).

By design, the COM, RLM and SZM are not able to characterize such a distribution, due to the pixel texture spacial localization absence. Then, in order to characterize such radial textures, we propose a descriptor named gray level Distance-to-border Zone Matrix (DZM), denoted $\mathcal{G}\mathcal{D}_f$. The new statistical $\mathcal{G}\mathcal{D}_f^{N_k}(d_n, g_m)$ matrix element yields the number of

intensity zones g_m at a d_n distance further to the E^c space support border. This distance is the shortest Euclidean distance between the flat zone and the shape border (see example of Fig. 3).

2	2	2	4	4	4	4	Level g_m	Distance zone, d_n			
2	1	1	4	4	1	1		0	1	2	3
3	1	2	2	2	1	4	1	3	1	0	1
3	4	4	1	2	1	4	2	2	0	1	0
3	4	4	4	3	3	4	3	1	1	0	0
2	2	2	3	3	3	1	4	3	1	0	0
1	1	4	4	4	1	1					

Level g_m	Size zone, s_n					
	1	2	3	4	5	6
1	1	1	2	1	0	0
2	0	0	1	2	0	0
3	0	0	1	0	1	0
4	0	0	2	0	1	1

Fig. 3. Four gray levels texture example, where each flat zone is valued with 4-connectivity and distance to the border E^c (top left), with the resulting DZM (top right) and SZM (bottom).

In practice, to accelerate the computation time, the distance function is computed for the whole texture support space:

$$D(\mathbf{x}, E) = \inf\{d(\mathbf{x}, \mathbf{y}), \mathbf{y} \in E^c\}$$

where $d(\mathbf{x}, \mathbf{y})$ is typically obtained using a discrete metric approximation to the Euclidean distance (Chamfer or Montanary). Then, for each region $R_j[f]$, the corresponding distance value is obtained as its smallest value in the distance map:

$$d_j = \inf\{D(\mathbf{z}, E), \mathbf{z} \in R_j[f]\}$$

Remarks:

- Generalize this matrix to construct a Multiple gray levels Distance-to-border Zone Matrix $\widetilde{\mathcal{G}}\mathcal{D}_f$ is obvious:

$$\widetilde{\mathcal{G}}\mathcal{D}_f(d_n, g_m) = \sum_{k=1}^8 w_k \mathcal{G}\mathcal{D}_f^{N_k}(d_n, g_m)$$

- The distance is the shortest Euclidean distance between the shape flat zone and the border, but we can use a variant using the distance from the flat zone barycenter (more representative for a global approach) to the border, in order to handle long or large flat zones which touch the border.
- Flat zones sizes are not taking into account and we can suppose that, in many applications, large zones are more significant than small ones. Hence we can create a weighted matrix in which $\mathcal{G}\mathcal{D}_f(d_n, g_m)$ yields the zones sizes sum of intensity g_m at a distance of d_n from the border.

VI. ORIENTATION LENGTH ZONE MATRIX

The SZM and DZM are statistical descriptors assuming that the texture is composed of a randomized homogeneous zones non periodic tiling, each one described by its intensity value and its size/distance-to-border. This principle, which is appropriate to describe intensity-dependent homogeneous vs. heterogeneous textures [34], is not compatible with other

structured textures kinds; in particular, with fibrous textures. Let us consider the given microtubule network example in Fig. 4 (a), which is another studied case in Section VII. Similar texture images can be found in other natural objects such as wood, carbon, wool, all fibrous materials, etc. On the one hand, we observe that the fiber intensity is not an important feature. Fibers can mainly be described by their length, width variation, orientation and tortuosity [36]. We assume here that our fibrous textures are thin fibers randomized network of limited width variation and low tortuosity. On the other hand, the fibrous textures segmentation by flat zones is without interest for construction of descriptors which characterize fiber network morphological properties.

Hence we offer to use the Local Radon Transform (LRT), as discussed in [37], to segment individual fibers. The LRT uses an orientated Gaussian derivative kernel, rotated at different angles and adapted via a maximization procedure to the various texture directions (see Fig. 4). Other filtering oriented techniques could have been used, but the LRT intrinsically provides a fast (implementation using FFT) and robust (regularization by the smallest Gaussian) result. Now, the connected flat zones computation (two neighboring points belonging to the same zone if they have the same dominant orientation value) produces a network segmentation in J linear segments which roughly represents each fiber $F_j[f]$, but which does not cover the whole support space, i.e., $\cup_{j=1}^J F_j[f] \neq E$; in addition, in the fibers crossing points, one orientation is arbitrary favored over the others. We solved this last drawback by considering separately the connected components of each orientation and by reconnecting them with a small morphological closing: for each orientation, concerning fibers are isolated and connected with an unitary oriented structuring element. To improve the descriptor robustness, the regions associated to very small fibers can be rejected. Once the individual fibers segmentation is available, each fiber can be described by its length l_j (computed as its geodesic diameter) and by its orientation θ_j (main axis orientation, computed by PCA [38]). Using these two parameters, we propose to characterize a fibrous texture f by a new statistical matrix named the Orientation Length Zone Matrix (OLZM), $\mathcal{OL}_f(\theta_n, l_m)$, which yields the “fibers” number of orientation θ_n and length l_m . The OLZM rows number depends on the orientation space discretization degree (which is selected in the LRT computation) and the columns number equals to the longest texture fiber.

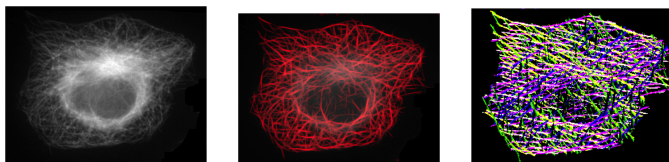


Fig. 4. The original microtubule network (left); an enhanced network image (with LRT) giving at each pixel \mathbf{x} the image processed value with a kernel orientated at angle θ_i which produces the maximal intensity (middle); an orientation map gives at each \mathbf{x} , the corresponding maximal response θ_i represented by a different color label (right).

To deal with the rotation invariance problem, we propose the

following alternative solutions. If the texture f is a segmented object, for instance a cell such as the ones in our case study, each zone orientation is given with respect to the coordinates system associated to the object main axis (computed by PCA). If the texture f is not bounded in the image, the main axis can be replaced by the texture full zone orientation average.

VII. APPLICATION TO CELL CLASSIFICATION

A. Cells pattern

In order to assert our advanced statistical matrices texture description performances, we compare their results with the ones obtained with the powerful morphological pattern spectrum (PS), providing a texture morphological multi-scale representation. Pattern spectrum is based upon the granulometry notion. A granulometry (resp. anti-granulometry) is an image objects size distribution study [39], [40]. Formally, for the discrete case, a granulometry (resp. anti-granulometry) is an opening family $\Gamma = (\gamma_{B_n})_{n \geq 0}$ (resp. closings $\Phi = (\varphi_{B_n})_{n \geq 0}$) that depends on an integer positive parameter n (which expresses a size factor). We remind that the f image opening using the structuring element B of size n is obtained by concatenation: an erosion with B_n followed by a dilation with the same structuring element [39], i.e., $\gamma_{B_n}(f) = \delta_{B_n}(\varepsilon_{B_n}(f))$. The closing is obtained by reversing the operators order, i.e., $\varphi_{B_n}(f) = \varepsilon_{B_n}(\delta_{B_n}(f))$. The f image granulometric analysis with respect to Γ consists in evaluating each opening of size n with a measurement, typically the opened image integral, i.e., $\sum_{\mathbf{x} \in E} \gamma_{B_n}(f)(\mathbf{x}) dx$. The granulometric curve, or pattern spectrum $PS(f, n)$ [40] of f with respect to Γ and Φ , is defined by the following normalized mapping:

$$PS(f, n) = \frac{1}{\sum_{\mathbf{x} \in E} f(\mathbf{x}) dx} \begin{cases} \sum_{\mathbf{x} \in E} \gamma_{B_n}(f)(\mathbf{x}) dx - \\ \sum_{\mathbf{x} \in E} \gamma_{B_{n+1}}(f)(\mathbf{x}) dx, \\ \text{for } n \geq 0 \\ \sum_{\mathbf{x} \in E} \varphi_{B_{|n|}}(f)(\mathbf{x}) dx - \\ \sum_{\mathbf{x} \in E} \varphi_{B_{|n|-1}}(f)(\mathbf{x}) dx, \\ \text{for } n \leq -1 \end{cases}$$

The pattern spectrum value for each size n corresponds to structures measurement of size n and is a probability density function (i.e. a histogram): a peak or mode in PS at a given scale n indicates the presence of many image structures of this scale (or size). Granulometric size distributions can be used as descriptors for texture classification. We use both granulometry and anti-granulometry to characterize bright and dark structures; that is, cell speckles in this application.

Tables I, II and III present results, on the *ICPR 2012 Hep-2 cells classification contest* dataset, for both characterization techniques families previously described, and for three classifiers. For SZM, the gray levels image number is reduced to 64 (empirically estimated). For the PS, structuring elements sizes are from $n = 1$ to $n = 13$ with a 2 size step, in order to detect small to big speckles, where the structuring element B is a discrete disk unit.

Considered separately, every technique is efficient (predictions upper than 90%, except for the “centromere” class) for logistic regression and quite perfect results for random forest and neural network. However some rare mistakes remain. A

	COM	RLM	SZM	MSZM	PS
Centromere	81.38	85.67	88.88	84.41	92.69
Coarse speckles	93.87	97.38	98.2	95.42	91.91
Cytoplasmic	98.26	99.1	99.1	99.39	97.06
Fine speckles	88.2	85.41	97.56	90.75	90.75
Homogeneous	93.7	93.78	97.81	95.18	93.61
Nucleolar	90.47	91.6	93.46	92.41	92.08

TABLE I

CLASSIFICATION RESULTS (IN PERCENTAGE) WITH LOGISTIC REGRESSION

	COM	RLM	SZM	MSZM	PS
Centromere	95.61	96.59	97.86	98.15	99.31
Coarse speckles	98.04	99.26	99.59	99.51	99.51
Cytoplasmic	99.17	99.77	100	100	100
Fine speckles	94.81	97.52	98.48	99.44	99.36
Homogeneous	96.93	97.37	98.42	98.86	98.04
Nucleolar	98.22	98.62	99.53	99.6	98.95

TABLE II

CLASSIFICATION RESULTS (IN PERCENTAGE) WITH RANDOM FOREST.

	COM	RLM	SZM	MSZM	PS
Centromere	95.66	93.47	94.93	96.59	97.95
Coarse speckles	99.02	98.69	99.02	99.26	99.26
Cytoplasmic	99.7	99.32	99.85	99.85	99.85
Fine speckles	96.04	93.78	96.33	97.29	99.04
Homogeneous	97.9	93.17	97.9	97.99	98.25
Nucleolar	98.27	98.95	98.71	99.27	98.71

TABLE III

CLASSIFICATION RESULTS (IN PERCENTAGE) WITH NEURAL NETWORK.

systematic study revealed us that they are different following to the characterization techniques and classifiers. Therefore we considered a final weighted average probability, where the weights correspond to the empirical predictions given on the Tables. This classifiers combination improves our results, leading to a perfect 100% prediction for all classes.

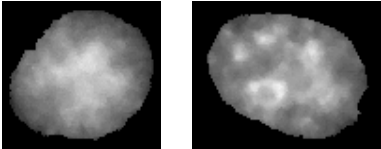
B. Cell division assays

The development of the present statistical texture descriptors was motivated by an application in cell-based assays for phenotypic screening, which consists in the use of multi-parametric and high resolution imaging techniques to characterize and select innovative compounds and/or protein targets involved in cell division. More precisely, the aim is to build a cell phase classifier by the DNA structure analysis (using a nuclear chromatin marker) and the microtubule network organization (using a cytoskeleton marker). We disposed of the RAMIS dataset made with 317 cells which were extensively annotated by experts. Annotation includes the mitosis phase, with approximately 50 other labels providing relevant information for phase classification. For this experimental results part, logistic regression is considered as classifier, with *One Class Classifiers* for each class, and validation by *Leave One Out* (k-fold validation with k equals

to the number of instances).

Among all available labels, we here focus on:

- 1) *DNA Texture Homogeneity*, containing two classes *Homogeneous* and *Heterogeneous*. Fig. 5 shows that MSZM provides comparable result than the best original SZM. This property can be observed in Fig. 7 too for radial distribution. Confidence intervals sizes are equal and models probabilities are lower than 0.001.



Matrix	Prediction	Confidence Interval
PS	91.6	—
$RLM_f(s_n, g_m)$	89.4	[87.6, 91, 2]
$GS_f^8(s_n, g_m)$	89.3	—
$GS_f^{16}(s_n, g_m)$	91.1	—
$GS_f^{32}(s_n, g_m)$	92.7	[89.9, 95.6]
$GS_f^{64}(s_n, g_m)$	91.6	—
$\tilde{G}S_f(s_n, g_m)$	92.7	[89.9, 95.6]

Fig. 5. Examples of homogeneous texture (left) vs. heterogeneous texture (right) nuclei; and results of homogeneity classification (in percentage) using second-order moments of GLZSM.

We can observe in Fig.6 the *Receiver Operator Characteristic* curves (ROC, the true positive rate / sensitivity is plotted in function of the false positive rate / 100-Specificity for different cut-off points of a parameter) and the *Area Under ROC* (AUR, the closer to 1 the better), in order to estimate the results sensitivity/specificity. The best SZM specific gray levels number obtains the best area under curve, the MSZM has a comparable result, even better than RLM one. Moreover the probabilities distribution is better for SZM and MSZM than RLM. MSZM has the best repartition close to the extremities (so really few ambiguous cases), but SZM has the strong errors lowest number (misclassified nuclei with strong probability). These results demonstrate the power and usefulness of such a multiple gray levels version. Moreover, we have empirically observed that MSZM has generally better or comparable results than SZM. This can be observed in both table II and table III.

- 2) *DNA Masses Texture*, which is the DNA “distribution” and contains four classes (*Beads*, *Slightly Condensed*, *Condensed* and *Highly Condensed*). Fig. 7 shows that MDZM provides a better radial distribution description than MSZM. However, there is an exception for class “Beads”, which is composed of textures without radial repartition (so random or homogeneous texture). Thus MSZM is well adapted and provides better results.
- 3) *Microtubules Network Organization*, which contains three classes (*Well Organized*, *ReOrganized* and *Other*). Fig. 8 shows OLZM really satisfying results and this

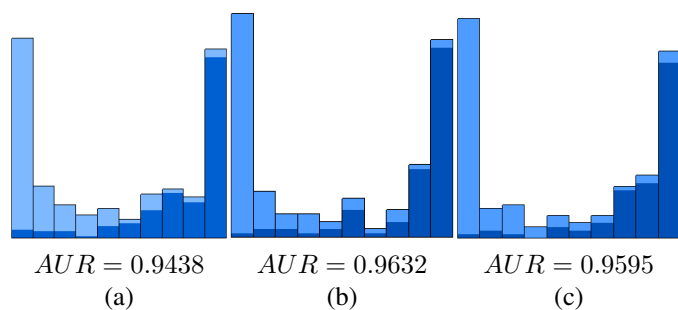


Fig. 6. The probabilities histograms and the area under roc curve (AUC), given by each model: RLM (a), SZM (b) and MSZM (c). In dark blue the nuclei with heterogeneous texture and the probability 1 (resp. 0) is on the right (resp. left). The closer to 1 (resp. 0) the probability, the more heterogeneous (resp. homogeneous) the texture.

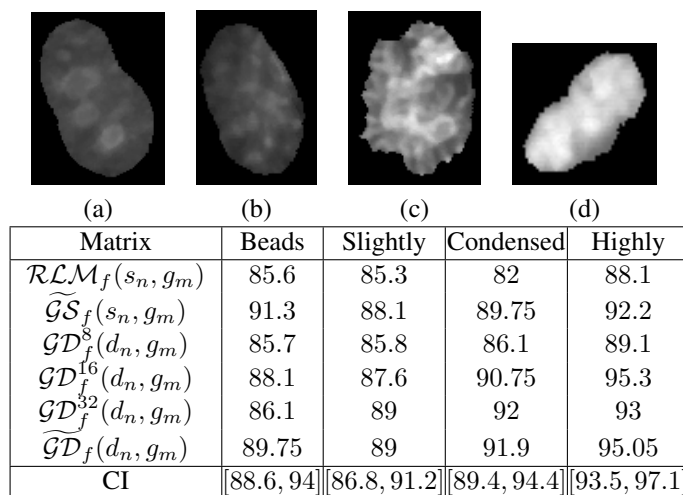


Fig. 7. Nuclei examples with different textures (associated here to the chromatin condensation): beads (a), slightly (b), condensed (c), highly (d); classification results (in percentage) and best results confidence intervals.

is the first method proposed to figure out this problem. However, these results are slightly lower than other results presented in this paper, due to the organization similarity among texture classes.

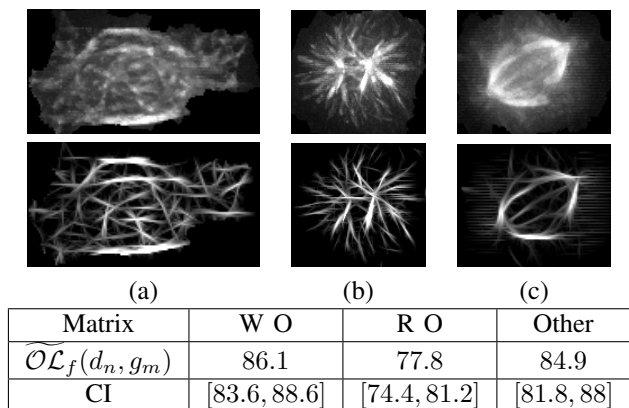


Fig. 8. Examples of microtubule network normal/enhanced with different organization: Well Organized WO (a), ReOrganized RO (b), Other (c); the results of classification (in percentage) and the confidence intervals CI.

VIII. CONCLUSION AND PERSPECTIVES

In this paper, the problem of cell characterization and classification was addressed. More precisely, we focused on specific and relevant parts of cell characterization particularly difficult during mitosis. To figure out these problems, we designed new advanced statistical matrices based on the Gray Level Size Zone Matrix. First, a multiple gray levels version which uses more information about the texture thanks to a complete gray level decomposition. It provides at least comparable results to the best original SZM on both applications, uses one less parameter, but requires more computations. In addition, two new versions, which use radial distribution and length/orientation of flat zones, in order to characterize specific DNA and microtubules aspects. These matrices showed their power and efficiency for quantitative cell analysis, and can be applied to other specific problems of texture characterization.

ACKNOWLEDGMENT

This work was part of RAMIS project (2007-2010) funded by the General Directorate for Competitiveness, Industry and Services of the French Ministry for the Economy, Industry and Employment. The authors thank Chantal Etievant, Delphine Reberieux, Vanessa Tillement and Michel Wright from Pierre Fabre pharmaceutical company for providing the annotated cell images, and Ronan Danno from ADCIS company for the annotation software.

REFERENCES

- [1] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biology*, vol. 7, p. R100, 2006. [Online]. Available: <http://www.cellprofiler.org/index.shtml>
- [2] P. Perner, H. Perner, and B. Müller, "Texture classification based on random sets and its application to hep-2 cells," in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 2002, pp. 406–411.
- [3] B. Newmann and T. Walker, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes," *Nature*, vol. 464, no. 7289, pp. 721–7, 2012.
- [4] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari, "Shape and texture indexes: Application to cell nuclei classification," *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, vol. 27, no. 1, 2013. [Online]. Available: <http://www.worldscientific.com/doi/abs/10.1142/S0218001413570024>
- [5] G. Thibault, J. Angulo, and F. Meyer, "Advanced statistical matrices for texture characterization: Application to dna chromatin and microtubule network classification," in *IEEE International Conference on Image Processing (ICIP)*, September 2011, pp. 53–56.
- [6] P. Foggia, G. Percannella, P. Soda, and M. Vento, "Benchmarking hep-2 cells classification methods," *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, June 2013.
- [7] J. Angulo, D. Reberieux, G. Thibault, C. Etievant, and F. Meyer, "Self-normalization of cell images in multifocus quantitative fluorescence," in *International Congress of Stereology*, Beijing, China, October 2011, p. 4.
- [8] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, December 1998.
- [9] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1145.
- [10] M. Stone, "Cross-validated choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. 36, no. 2, pp. 111–147, 1974.

- [11] H. Martens and P. Dardenne, "Validation and verification of regression in small data sets," *Chemometrics and intelligent laboratory systems*, vol. 44, no. 1-2, pp. 99-121, 1998.
- [12] J. Berkson, "Application of the logistic function to bio-assay," *Journal of the American Statistical Association*, vol. 39, pp. 357-365, 1944.
- [13] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*. Toronto: John Wiley & Sons, 1989.
- [14] McFadden, *Conditional logit analysis of qualitative choice behaviour*, P. Z. (ed.), Ed., 1974.
- [15] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.
- [16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. CRC Press, 1984.
- [18] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
- [19] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural computation*, vol. 9, no. 7, pp. 1545-1588, 1997.
- [20] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115-133, 1943.
- [21] I. A. Liu, J. Ghosh, and C. Martin, "Generative oversampling for mining imbalanced datasets," in *International Conference on Data Mining*, R. Stahlbock, S. F. Crone, and S. Lessmann, Eds. CSREA Press, 2007, pp. 66-72.
- [22] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory under-sampling for class-imbalance learning," in *IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, December 2006, pp. 965-969.
- [23] S. Marcellin, D.-A. Zighed, and G. Ritschard, "An asymmetric entropy measure for decision trees," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2006, pp. 1292-1299.
- [24] N. Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies," in *AAAI Workshop on Learning from Imbalanced Data Sets*. AAAI Press, 2000, pp. 10-15. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.1396>
- [25] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610-621, 1973.
- [26] O. S. Al-Kadi, "Texture measures combination for improved meningioma classification of histopathological images," *Pattern Recognition*, vol. 43, pp. 2043-2053, May 2010.
- [27] Q. Ji, J. Engel, and E. Craine, "Texture analysis for classification of cervix lesions," *IEEE Transactions on Medical Imaging*, vol. 19, no. 11, pp. 1144-1149, November 2000.
- [28] J. Wang, X. Zhou, P. L. Bradley, S.-F. Chang, N. Perrimon, and S. T. Wong, "Cellular phenotype recognition for high-content rnai genome-wide screening," *Journal of Biomolecular Screening*, vol. 13, no. 1, pp. 29-39, January 2008.
- [29] M. Galloway, "Texture analysis using grey level run lengths," *Computer Graphics Image Processing*, vol. 4, pp. 172-179, July 1975.
- [30] A. Chu, C. Sehgal, and J. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognition Letters*, vol. 11, no. 6, pp. 415-419, 1990.
- [31] B. Dasarathy and E. Holder, "Image characterizations based on joint gray level run length distributions," *Pattern Recognition Letters*, vol. 12, no. 8, pp. 497-502, 1991.
- [32] L. S. Davis, M. Clearman, and J. Aggarwal, "A comparative texture classification study based on generalized cooccurrence matrices," in *IEEE Conference on Decision and Control*, Miami FL, December 1979.
- [33] H. Schulerud, J. M. Carstensen, and H. Danielsen, "Multiresolution texture analysis of four classes of mice liver cells using different cell cluster representations," in *The 9th Scandinavian Conference on Image Analysis*, Uppsala, Sweden, 1995, pp. 121-129. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.3125>
- [34] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari, "Texture indexes and gray level size zone matrix. application to cell nuclei classification," in *Pattern Recognition and Information Processing (PRIP)*, Minsk, Belarus, May 2009, pp. 140-145.
- [35] F. Meyer, "The levelings," in *Mathematical Morphology and its Applications to Image and Signal Processing*, Heijmans and R. Eds, Eds., Kluwer, 1998, pp. 199-206.
- [36] L. Decker, D. Jeulin, and I. Tovená, "3d morphological analysis of the connectivity of a porous medium," *Acta Stereologica*, vol. 17, no. 1, pp. 107-112, 1998.
- [37] M. Krause, R. Alles, B. Burgeth, and J. Weickert, "Retinal vessel detection via second derivative of local radon transform," Department of Mathematics, Saarland University, Tech. Rep. 212, June 2008.
- [38] H. Hotelling, "Analysis of a complex of statistical variables with principal components," in *Journal of Educational Psychology*, 1933.
- [39] J. Serra, *Image Analysis and Mathematical Morphology*. London: Academic Press, 1982, vol. 1.
- [40] P. Maragos, "Pattern spectrum and multiscale shape representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 701-716, July 1989.



Dr. Guillaume THIBAUT received a degree in computer science, with a M.S. thesis in image analysis in 2005, and the Ph.D. degree in image processing and pattern recognition from the University of Aix-Marseille II, France, in 2009. He was former a research engineer in the Center of Mathematical Morphology (CMM) at MINES-ParisTech and is currently a research engineer in the Oregon Health & Science University.

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



Dr. Jesús Angulo was born in Cuenca, Spain, in 1975. He received a degree in Telecommunications Engineering from Polytechnical University of Valencia, Spain, in 1999, with a Master Thesis on Image and Video Processing. He obtained his PhD in Mathematical Morphology and Image Processing, from the Ecole des Mines de Paris (France), in 2003, under the guidance of Prof. Jean Serra. He is currently a permanent researcher (Chargé de Recherche) in the Center of Mathematical Morphology (Department of Mathematics and Systems) at MINES ParisTech.

His research interests are in the areas of multivariate image processing (colour, hyper/multi-spectral, temporal series, polarimetric, tensor imaging) and mathematical morphology (filtering, segmentation, shape and texture analysis, stochastic and geometry approaches, PDE approaches), and their application to the development of theoretically-sound and high-performance algorithms and software in the fields of Biomedicine/Biotechnology, Remote Sensing and Industrial Vision.



Pr. Fernand Meyer got an engineer degree from the Ecole des Mines de Paris in 1975. He works since 1975 at the Centre de Morphologie Mathématique (CMM) of the Ecole des Mines de Paris, where he is currently director. His first research area was "Early and automatic detection of cervical cancer on cytological smears", subject of his PhD thesis, obtained in 1979. He participated actively to the development of mathematical morphology: particle reconstruction, top-hat transform, the morphological segmentation paradigm based on the watershed transform and markers, the theory of digital skeleton, the introduction of hierarchical queues for high speed watershed computations, morphological interpolations, the theory of levelings, multi-scale segmentation.