Efficient Statistical/Morphological Cell Texture Characterization and Classification

Guillaume Thibault
CMM, MINES-ParisTech
Guillaume.Thibault@MINES-ParisTech.fr

Jesus Angulo
CMM, MINES-ParisTech
Jesus.Angulo@MINES-ParisTech.fr

Abstract

This paper presents the different steps for an automatic fluorescence-labelled cell classification method. First a data features study is discussed in order to describe cell texture by means of morphological and statistical texture descriptors. Then, results on supervised classification using logistic regression, random forest and neural networks, for both morphological and statistical descriptors, is presented. We propose a final consolidated classifier based on a weighted probability for each class, where the weights are given by the empirical classification performances. The method is evaluated on ICPR'12 HEp-2 dataset contest.

1. Introduction

Current technologies of parallel cells growing in multi-well plates (or in other supports as cell on chip) and fluorescent labeling of proteins of interest (immuno-fluorescence with antibodies, GFP-tagged proteins), together with image capture by automated microscopy and subsequent cell image analysis. This is of interest for the discovery of new cellular biology mechanisms (i.e., using siRNA), new pharmaceuticals (i.e., mass screening of potential active molecules) or for the development of new tests for diagnostic/prognostic, for toxicology tests (i.e., evaluation of different compounds at different concentrations). The larger number of cells acquired, the sounder analysis is obtained. Currently, most of this processing is manual, being time consuming and involving variability of results according to the expert (inter-observer variability). To achieve a robust high throughput system which will be able to automatically analyze thousands of cell images without needing a manual interaction, and in particular supervised cell classification is one the key point for such systems [10]. Cell classification is a classical task in pattern recognition [4, 11].

This paper presents a method to classify automatically cells with an excellent accuracy. The approach is based on using two different families of descriptors and three different classifiers. That involves a consolidation of the description/classification which notably improve the results. The cell dataset considered to illustrate our approach was acquired by indirect immunofluorescence (IIF) and provided by the ICPR 2012 *HEp-2 Cells Classification* contest[1]. It contains 1457 cells divided into 6 classes:

- **Centromere** (388), several discrete speckles (40-60) distributed throughout the interphase nuclei and characteristically found in the condensed nuclear chromatin during mitosis as a bar of closely associated speckles.
- **Coarse** (239) or **fine** (225) **speckled**, granular nuclear staining of the interphase cell nuclei.
- **Cytoplasmic** (128), fine fluorescent fibers running the length of the cell.
- **Homogeneous** (345), diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells.
- **Nucleolar** (257), large coarse speckled staining within the nucleus, less than six in number per cell.

The first step towards classification is an appropriate feature extraction. Section 2 presents the two families of extracted features to describe cell patterns, based on statistical matrices (2.2) and pattern spectrum (2.3). Next, Section 3 discusses the classification strategy based on three supervised algorithms. Results on ICPR 2012 contest are given in Section 4.

[1] http://mivia.unisa.it/hep2contest/index.shtml

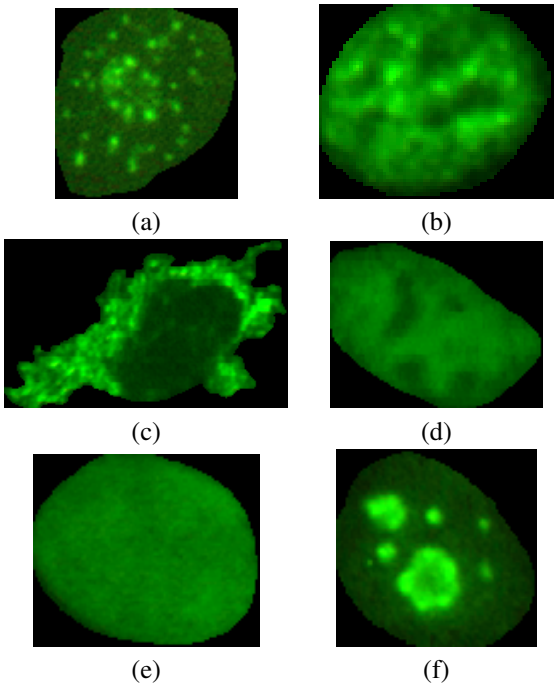


Figure 1. Examples of typical cells for each category: (a) centromeres, (b) coarse speckle, (c) cytoplasmic, (d) fine speckle, (e) homogeneous and (f) nucleolar.

2. Cell Feature Extraction

Let $f(\mathbf{x}) : E \rightarrow \mathcal{T}$ be a gray-level image, where $E \subset \mathbf{Z}^2$ is the space pixels $\mathbf{x} \in E$ and the image intensities are discrete values which range in a closed set $\mathcal{T} = \{t_1, t_2, ..., t_N\}$, $\Delta t = t_{i+1} - t_i$, e.g., for a 8 bits image we have $t_1 = 1$, $N = 256$ and $\Delta t = 1$. Let us assume also that image f is segmented into its J flat zones (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 $s(j) = |R_j[f]|$. Hence, we consider that each zone $R_j[f]$ has associated a constant gray-level intensity $g(j)$.

On the other hand, mathematical morphology operators for gray level images [1] can be applied on image f. In particular, opening defined by the combination of an erosion followed by a dilation, i.e., $\gamma_B(f) = \delta_B(\varepsilon_B(f))$, and the dual closing, i.e., $\varphi_B(f) = \varepsilon_B(\delta_B(f))$, are operators which extract respectively bright and dark image structures according to the shape/size of the structuring element B.

2.1. Cell morphology variability

When we observe cells in figure 1, it appears clearly that differences between cell classes are mainly based on the presence and distribution (numbers, sizes and intensity) of bright/dark structures such as centromeres, speckles and nucleolus: numbers, sizes (areas) and colors. Homogeneous cells do not contain such structures, coarse and fine classes have dark speckles of various sizes, centromeres are brighter than speckles, etc. Even cytoplasm texture is rough with strong intensity variation. In conclusion, it seems natural to use characterization methods that provide an analysis of texture and more specifically methods able to deal with bright/dark speckle-like structure description.

2.2. Gray level size zone matrix (GLSZM)

The GLSZM matrix [13, 14] is the starting point of Thibault's matrices. The GLSZM of a texture image f, denoted $\mathcal{GS}_f(s_n, g_m)$, provides a statistical representation by the estimation of a bivariate conditional probability density function of the image distribution values. It is calculated according to the pioneering Run Length Matrix principle [6]: the value of the matrix $\mathcal{GS}_f(s_n, g_m)$ is equal to the number of zones of size s_n and of gray level g_m. The resulting matrix has a fixed number of lines equal to N (the number of gray levels) and a dynamic number of columns (determined by the size of the largest zone as well as the size quantization). The more homogeneous the texture, the wider and flatter the matrix. More precisely, we can calculate all the second-order moments of $\mathcal{GS}_f(s_n, g_m)$ as compact texture features [5]. Figure 2 shows an example of the calculation of such a matrix.

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

(a)

Level g_m	*Size zone*, s_n 1	2	3
1	2	1	0
2	1	0	1
3	0	0	1
4	2	0	1

(b)

Figure 2. Example of the GLSZM filling for an image texture of size 4×4 with 4 gray levels.

GLSZM does not require calculations in several directions, contrary to Run Length Matrix (RLM) [6] and Co-occurrences Matrix (COM) [7]. RLM and COM are appropriate for periodic textures whereas the GLSZM is typically adapted to describe heterogeneous non periodic textures. In addition, due to the intrinsic

segmentation, texture description in GLSZM is more regional than the point-wise-based representation of COM. However, it has been empirically proved that the degree of gray level quantization still has an important impact on the texture classification performance. For a general application it is usually required to test several gray level quantization in order to find the optimal one with respect to a training dataset.

GLSZM is particularly well adapted to the analysis and characterization of cell texture associated to speckle-like structures: it is wide and non null values are concentrated around an intensity value for homogeneous cells; wide with non null dark values due to a large zone for cytoplasm; and so on for the other classes.

2.3. Pattern spectrum (PS)

A granulometry (resp. anti-granulometry) is the study of the size distribution of the objects of an image [8, 12]. Formally, for the discrete case, a granulometry (resp. anti-granulometry) is a family of openings $\Gamma = (\gamma_n)_{n\geq 0}$ (resp. closings $\Phi = (\varphi_n)_{n\geq 0}$) that depends on a positive parameter n, which expresses a size factor for a fixed structuring element B. The granulometric analysis of an image f with respect to Γ consists in evaluating each opening of size n with a measurement: $\int \gamma_n(f)$. The granulometric curve, or pattern spectrum $PS(f,n)$ [8] of f with respect to Γ and Φ, is defined by the following normalized mapping:

$$PS(f,n) = \frac{1}{\int f} \begin{cases} \int \gamma_n(f) - \int \gamma_{n+1}(f), for\ n \geq 0 \\ \int \varphi_{|n|}(f) - \int \varphi_{|n|-1}(f), for\ n \leq -1 \end{cases}$$

The value of pattern spectrum for each size n corresponds to a measurement of structures 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 [1]. We use both granulometry and anti-granulometry to characterize speacklelike bright and dark cell structures, and the shape of B is a discrete disk.

3. Cell classification

There are several alternatives to deal with this scenario of 6 classes. Naturally, we can consider to classify cells according to an approach "*one against all*" classifier: each class has to be modeled in the feature space; however due to the number of classes, the boundaries between them are not easily separated. Therefore, we decided to work on an approach "*one class classifier*": for a given class C, all individuals which do not belong to C are regrouped together under the same label in order to reduce the problem to 2 classes (or binary classification problem). In such a case, the resulting working set is imbalanced. To solve this issue, we considered an over-sampling method [15]: elements of the minority class are duplicated until the creation of balanced classes. The global process (imbalance, over-sampling, classifier learning) is then realized for each class.
We used three different classifiers based on well-known supervised classification techniques:

- Logistic Regression [2] (RL) is a linear regression function particularly well adapted to binary classification problems, and usually preferred to more complex methods in order to avoid root learning.
- Random Forest [3] (RF) is one of the last advanced techniques in the aggregation of classification trees, and one of the most powerful in the current state-of-the-art.
- Neural network [9] (NN) is non-linear technique where learning consist in minimizing the cost associated to the average squared error using gradient descent (back-propagation on multilayer perceptrons).

These three classifiers are used with both characterization techniques. Hence, for each class, we dispose of six probabilities of classification. In order to introduce a consolidated classification result, the final probability is computed by a weighted average of the six, where the weights are given by the empirical classification performances presented in Tables 1 and 2.

4. Results

Tables 1 and 2 present the results of classification for both characterization techniques previously described and the considered three classifiers. For GLSZM, the number of image gray-level is reduced to 32 (empirically estimated) and for PS, sizes of structuring elements are from 1 to 13 with step of size 2, in order to describe from small to large cell structures such as centromeres and speckles.

Considered separately, both GLSZM and PS provides a prediction upper than 90% (except for GLSZM on "Centromere" class, see figure 3) for logistic regression and almost perfect accuracy for random forest and neural network. For the remaining mistakes (mainly between classes "fine-speackle" and "homogeneous"), a study reveals that they are different according to the

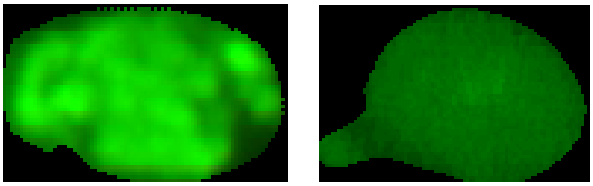
Classes	LR	RF	NN
Centromere	81.97	97.86	93.28
Coarse speckles	98.2	99.59	98.72
Cytoplasmic	99.1	100	99.43
Fine speckles	97.56	98.48	94.47
Homogeneous	97.81	98.42	96.78
Nucleolar	93.46	99.53	96.92

Table 1. Classification results with GLSZM (statistical descriptor).

Classes	LR	RF	NN
Centromere	92.69	99.51	97.62
Coarse speckles	91.91	99.51	98.55
Cytoplasmic	97.06	100	99.51
Fine speckles	90.75	99.36	97.41
Homogeneous	93.61	99.04	94.61
Nucleolar	92.08	98.95	97.18

Table 2. Classification results with PS (morphological descriptor).

characterization techniques and classifiers. There, our classification strategy of final weighted average probability is systematically right, and then the perfect result of 100% of prediction is reached for the 6 classes.



(a) (b)

Figure 3. Examples of misclassified cells with only logistic regression: homogeneous cells classified as fine speckle due to the presence of dark speackles.

5. Conclusions

This paper presents an application of statistical size zone matrix and morphological pattern spectrum to the automatic classification of cells. These methods provides powerful features who are then combined to three classifiers in order to provide an efficient classification of cells.

References

[1] *Morphological Image Analysis*. Springer-Verlag, 1999.

[2] J. Berkson. Application of the logistic function to bioassay. *Journal of the American Statistical Association*, 39:357–365, 1944.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] 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*, 7:R100, 2006.

[5] A. Chu, C. Sehgal, and J. Greenleaf. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11(6):415–419, 1990.

[6] M. Galloway. Texture analysis using grey level run lengths. *Computer Graphics Image Processing*, 4:172–179, July 1975.

[7] R. M. Haralick. Statistical and structural approaches to texture. In *Proceedings of the IEEE*, volume 67, pages 786–804, May 1979.

[8] P. Maragos. Pattern spectrum and multiscale shape representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):701–716, July 1989.

[9] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[10] B. Newmann and T. Walker. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–7, 2012.

[11] 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)*, volume 2, pages 406–411, 2002.

[12] J. Serra. *Image Analysis and Mathematical Morphology*, volume 1. Academic Press, London, 1982.

[13] 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)*, pages 53–56, September 2011.

[14] 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)*, pages 140–145, Minsk, Belarus, May 2009.

[15] G. M. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Artificial Intelligence Research*, 19:315–354, 2003.