# Image Analysis and Classification for High-Throughput Screening of Embryonic Stem Cells

**Laura Casalino, Pasqua D'Ambra, Mario R. Guarracino, Antonio Irpino, Lucia Maddalena, Francesco Maiorano, Gabriella Minchiotti, and Eduardo Jorge Patriarca**

**Abstract** Embryonic Stem Cells (ESCs) are of great interest for providing a resource to generate useful cell types for transplantation or novel therapeutic studies. However, molecular events controlling the unique ability of ESCs to self-renew as pluripotent cells or to differentiate producing somatic progeny have not been fully elucidated yet. In this context, the Colony Forming (CF) assay provides a simple, reliable, broadly applicable, and highly specific functional assay for quantifying undifferentiated pluripotent mouse ESCs (mESCs) with self-renewal potential. In this paper, we discuss first results obtained by developing and using automatic software tools, interfacing image processing modules with machine learning algorithms, for morphological analysis and classification of digital images of mESC colonies grown under standardized assay conditions. We believe that the combined use of CF assay and the software tool should enhance future elucidation of the mechanisms that regulate mESCs propagation, metastability, and early differentiation.

**Keywords** Classification • Colony assay • Imaging • Segmentation • Stem cells

L. Casalino • G. Minchiotti • E. Jorge Patriarca
Institute of Genetics and Biophysics "A. Buzzati-Traverso", CNR, Naples, Italy

P. D'Ambra (✉) • M.R. Guarracino • L. Maddalena • F. Maiorano
Institute for High-Performance Computing and Networking, CNR, Naples, Italy
e-mail: pasqua.dambra@na.icar.cnr.it

A. Irpino
Department of Political Science "J. Monnet", Second University of Naples, Caserta, Italy

# 1 Introduction

Application of image analysis and machine learning algorithms and tools to cell biology is a very active research field aimed to provide fast and objective methods for analyzing the large amount of images produced by modern high-throughput screening platforms available in biological research laboratories [3, 20, 22, 25].

In this work we describe first results related to the development of a multi-component software framework devoted to define automated morphological analysis and classification of Embryonic Stem Cells (ESCs) colonies. The colony-forming (CF) assay is widely used for monitoring the quality of ESC cultures as it currently offers the most sensitive and specific method to quantify the frequency of undifferentiated cells present in a culture. Moreover, it provides a reliable tool also for evaluating quantitative changes in pluripotent cell numbers, following manipulations that may affect the self-renewal and differentiation properties of the treated cells. In a clonogenic assay, under specific supportive conditions, pluripotent mouse embryonic stem cells (mESCs) form tridimensional round-shaped (*domed*) colonies. After the exposure to conditions that affect their metastability promoting an Epiblast-like phenotype [4] or induce differentiation, cells loose the ability to grow tridimensionally and form irregular and flattened (*flat*) colonies. The 4-day CF assay, by detecting the ability of mESCs to form domed or flat colonies, allows the composition of test cell populations to be quantified at the single cell level. However, to achieve acceptable statistical accuracy, a high number of cell colonies is required. In this regard, manual counting and classification are tedious, time-consuming, resource-intensive and subjective (operator-dependent). Therefore, the development of a reliable automated colony counter and classifier for such clonogenic assays would reduce time and resources required, while allowing greater statistical accuracy, standardization and reproducibility, thus offering the possibility for greater throughput over extended periods. To this aim, we proposed an experimental software tool which is able to automatically discriminate and quantify domed colonies, raising from undifferentiated self-renewing mESCs, and flat colonies, derived from undifferentiated Epiblast-like or differentiating cells. It is a multi-component framework interfacing different general-purpose software modules, implementing highly accurate algorithms for image pre-processing (spanning from region of interest identification to background removal), segmentation, and feature-based classification. Trained on untreated reference samples as well as on samples treated with reference compounds, the prototypal version of the software has been tested on a dataset of 40 microscopy images of single wells containing cells grown in different conditions. Comparison of the first results from automated versus manual colony segmentation and classification on randomly chosen images proved that the proposed tool is promising to be used as blind tool to support reliable analysis of molecular screening.

The paper is organized as follows. In Sect. 2 we describe the software components, both in terms of functionality and of main models and algorithms. In Sect. 3 we present first results and provide a quantitative performance analysis referring to usual performance metrics for segmentation and classification problems. Conclusions are drawn in Sect. 4.

## 2 Software System Components

According to the typical image analysis pipeline for High-Content Analysis [22], the proposed steps for automatically discriminating domed and flat colonies in microscopy images can be summarized as follows:

1. **Pre-processing**, to reduce image artifacts caused by imperfections in the image acquisition process.
2. **Segmentation**, to separate the cell colonies in each well image.
3. **Feature computation**, to provide numerical descriptors of each segmented colony.
4. **Classification**, to finally provide the discrimination and quantification of domed and flat colonies, based on the most discriminating features.

### 2.1 Pre-Processing

Input images are initially pre-processed, in order to allow an easier and more accurate segmentation of ESC colonies.

First of all, the well area is extracted by the whole image, in order to focus on the actual region of interest (ROI) in all subsequent steps, as exemplified in Fig. 1. This allows us to reduce not only the computational complexity of the entire procedure, but also the potentially misleading influence of non interesting image details (e.g., shadows and dark areas around the well, as shown in Fig. 1-a).
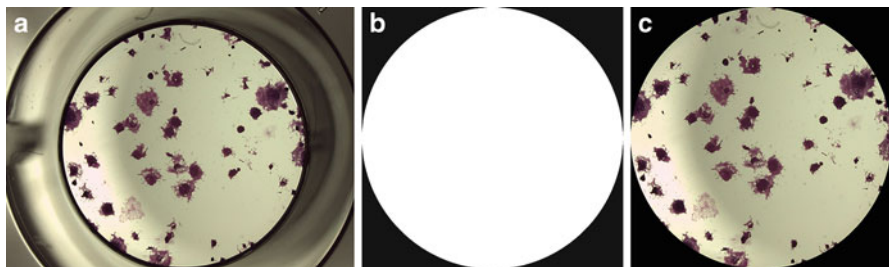


**Fig. 1** Extraction of the ROI: (**a**) original image; (**b**) circular ROI of the well; (**c**) obtained sub-image, which will be the input for subsequent steps (non interesting pixels shown in *black*)

Relying on the strong intensity discontinuities (edges) provided by the well borders, the circular ROI is accurately obtained through the Hough Transform [10], a powerful tool for the detection of parametric curves in images. It implements a voting process that maps image edge points into manifolds in an appropriately defined parameter space; peaks in this space correspond to the parameters of detected curves. Specifically, the Circle Hough Transform is designed to determine the parameters of a circle when a number of points that fall on its perimeter are known. A circle with radius $r$ and center $(x_0, y_0)$ can be described with the parametric equations

$$\begin{cases} x = x_0 + r\cos\theta \\ y = y_0 + r\sin\theta \end{cases}.$$

When the angle $\theta$ sweeps through the full 360 degree range, the points $(x, y)$ trace the perimeter of the circle. An image edge point $(x, y)$ is mapped to the 3D parameter space $(x_0, y_0, r)$, voting for all the circles that it could lie on. These votes are accumulated in a 3D array, whose maxima provide the parameters of most prominent circles. In the case of our single-well images, the accumulator maximum provides the well center pixel coordinates and the radius.

The next pre-processing step estimates and then removes the well background, in order to avoid the influence of eventual uneven illumination, as exemplified in Fig. 2. The goal is achieved through mathematical morphology operations [21], that, based on set theory, provide a tool to extract image components useful for the representation and description of region shape, and for pre- and post-processing of images.

Specifically, we relied on closing top-hat filtering, an operation that extracts small elements and details from a given image. Let $f$ be a grayscale image, and let $b$ be a grayscale structuring element. The closing top-hat transform of $f$ (sometimes called the bottom-hat transform or the black top-hat transform) is given by:
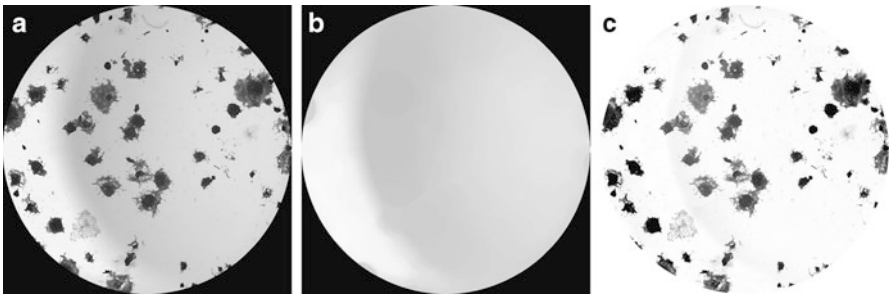
$$T_b(f) = f \bullet b - f,$$



**Fig. 2** Background estimation and removal: (**a**) input image (green color band of the image in Fig. 1-c); (**b**) estimated background; (**c**) obtained image, with background removed

where • indicates the closing operation. It returns an image containing the objects that are smaller than the structuring element and are darker than their surroundings. In the case of our images, morphological closing $f \bullet b$ using a structuring element $b$ having size greater than any expected cell colony removes all the colonies, thus providing a faithful approximation of the well background (see Fig. 2-b). The subsequent subtraction of the original image by the estimated background provides a uniformly illuminated image of the cell colonies, as shown in the normalized result reported in Fig. 2-c.

## 2.2 Segmentation

In order to partition each well image into its constituent objects (ESC colonies and background), we devised two different approaches. In the first approach, the image resulting from pre-processing is binarized through Otsu's method [17] (see Fig. 3-b) and then refined through binary morphological operations. Specifically, refinement involves morphological closing for removing small holes, hole-filling for removing internal holes, and removal of very small objects, i.e., small cell colonies or generic particles of no biological interest (see Fig. 3-c).

A second approach is based on a well-known variational model, for which we recently proposed efficient numerical solvers [9] with the final aim to develop software modules for modern high-performance environments [8]. In the following we briefly describe the model and the main features of the numerical approach and discuss the subsequent refinement of variational-based segmentation for application to the cell identification problem.

The image segmentation problem can be mathematically formulated in terms of a variational model, i.e., in terms of an energy minimization criterion. We look for a piecewise smooth function $u$ which approximates the original image function $f$, with $u$ discontinuous across a closed set $K$ that is included in the image domain $\Omega$ and represented by a suitable function $z$. In more details, let $\Omega \subset \Re^2$ be a bounded open
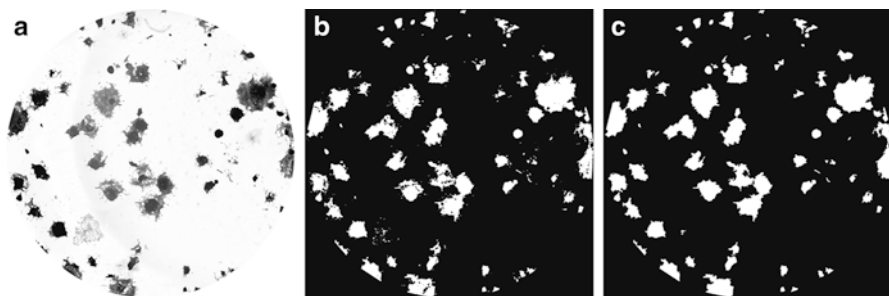


**Fig. 3** First approach to ESC colonies segmentation: (**a**) input image (the pre-processed image of Fig. 2-c); (**b**) binarized image; (**c**) refined binarized image

set and $f \in L^{\infty}(\Omega)$ the observed gray-level image. The problem can be described in terms of the minimization of the following functional:

$$E_{\epsilon}(u,z) = \int_{\Omega} (u-f)^2 dxdy + \beta \int_{\Omega} z^2 |\nabla u|^2 dxdy + \alpha \int_{\Omega} \left( \epsilon |\nabla z|^2 + \frac{(z-1)^2}{4\epsilon} \right) dxdy,$$

where $u \in C^1(\Omega \setminus K)$, $0 \le z \le 1$ is a function which controls $|\nabla u|$ and gives an approximate representation of the set $K$, $\alpha$ and $\beta$ are positive coefficients and $\epsilon$ is a positive sufficiently small parameter. The choice of the parameter set, especially of $\epsilon$, affects the accuracy in the detection of $K$ and it is a main drawback in the numerical solution of 1. Our approach for the solution of 1, known as the Ambrosio-Tortorelli model or phase-field approximation of the Mumford-Shah model [1], is based on the classical approach of Calculus of Variations which consists in writing the corresponding Euler-Lagrange equations, representing the necessary conditions for the minimizer. Euler-Lagrange equations for 1 are a non-linear system of elliptic partial differential equations coupled with Neumann conditions, which can be solved by finite-difference discretization. We proposed to apply a second-order finite-difference scheme and, starting from a block form of the resulting algebraic equations, we applied a first-order alternating minimization scheme based on the non-linear Gauss-Seidel method, accelerated by inner linear iterations. Our numerical results and comparisons with standard numerical techniques show the efficiency and the robustness of our solution approach for increasing sizes of cell colony images. Details on the numerical scheme and on the above results are discussed in [9]. Here, we only point out that results discussed in the following have been obtained, starting from the preprocessed image obtained by the green band of the original RGB image, when the parameters in 1 were set as follows: $\alpha = 1$, $\beta = 2$ and $\epsilon = 10^{-3}$.

The obtained function $z$ (see Fig. 4-a), also known as the edge set of the image, represents the set $K$ where the piecewise smooth function $u$ (also known as the restored image) is discontinuous and allow us to identify the borders of the cell colonies. Indeed, starting from the edge set, we applied subsequent refinements,
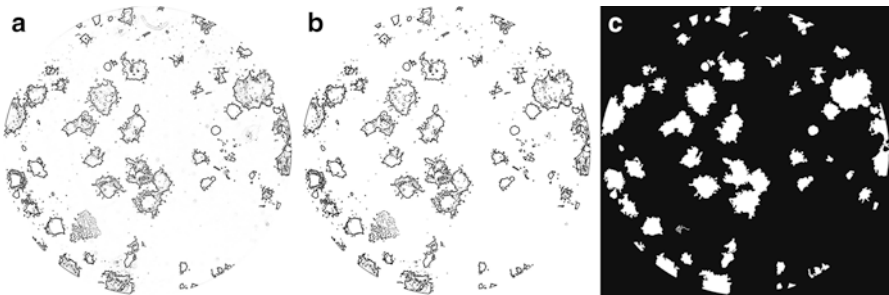


**Fig. 4** Second approach to ESC colonies segmentation: (**a**) edge set of the preprocessed image of Fig. 2-c; (**b**) binarized version of the edge set; (**c**) refined binarized image
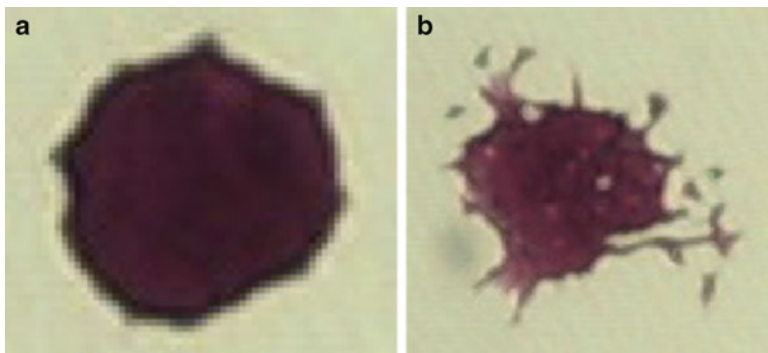
**Fig. 5** Examples of ESC colonies extracted from the image in Fig. 1: (**a**) a domed colony (size about 4 % of the original image); (**b**) a flat colony (size about 7 % of the original image)

based on binarization (see Fig. 4-b), hole-filling for removing internal holes, and finally removal of very small objects, analogous to those of the previously described approach, which leads to the cell colony segmentation (see Fig. 4-c).

## 2.3 Features Computation

In order to classify the ESC colonies and be able to differentiate between domed and flat ones, we designed a general approach to feature selection and classification, based only on distinctive feature estimates that can be computed by the available images.

The most distinctive features that allow the biologist to discriminate the two kinds of colonies concern their shape: domed colonies appear compact and rounded, as opposed to flat colonies that spread more or less throughout the well (see Fig. 5) This prior knowledge is exploited by using geometric models locally estimated for each colony in a well. Among the considered shape features, the most significative are those reported in Table 1. The *Area A(X)* provides the number of pixels of the object (colony) $X$ and the *Perimeter P(X)* is computed as the sum of the distances between consecutive pairs of boundary pixels of $X$. *Solidity S(X)* is given by the ratio of the area $A(X)$ of the object $X$ and the area $A(CH(X))$ of its convex hull [5]. Solidity ranges in [0,1], producing low values when the shape of the object shows many concavities (as in the flat colonies) and high values when the shape of the object shows few or zero concavities (as in the domed colonies). The irregularity of a contour is expressed through *Compactness C(X)* [7], given by the ratio between the area $A(X)$ of the object and $P^2(X)/4\pi$, the area of a circle having the same perimeter $P(X)$ as the object. Indeed, this feature, ranging in [0, 1], provides low values for scarcely compact shapes (as in the flat colonies) and high values for compact shapes (as in the domed colonies), reaching its maximum for the most

**Table 1** Some of the adopted shape (S) and texture (T) features and their values for the ESC colonies of Fig. 5

| Feature | Description | Value for Fig. 5-a | Value for Fig. 5-b |
|---|---|---|---|
| (S) Area | $A(X) = $ # pixels of object $X$ | 1512 | 3258 |
| (S) Perimeter | $P(X) = $ length of border of object $X$ | 153.09 | 496.54 |
| (S) Solidity | $S(X) = A(X)/A(CH(X))$ | 0.93 | 0.65 |
| (S) Compactness | $C(X) = 4\pi A(X)/P^2(X)$ | 0.81 | 0.18 |
| (T) Entropy | $H(X) = -\sum_i \sum_j M(i,j) \log_2 M(i,j)$ | 6.75 | 6.96 |
| (T) Contrast | $CN(X) = \sum_i \sum_j (i-j)^2 M(i,j)$ | 0.28 | 0.30 |
| (T) Energy | $E(X) = \sum_i \sum_j M(i,j)^2$ | 0.37 | 0.14 |
| (T) Correlation | $CR(X) = \frac{\sum_i \sum_j (i-\mu_i)(j-\mu_j)M(i,j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}$ | 0.96 | 0.95 |

compact shape: the circle. Other adopted shape features are: *Eccentricity*, that is the eccentricity of the minimum area ellipse including the colony image; *Min* and *Max axis length*, that are the minimum and the maximum axis length of the ellipse containing the colony image, respectively; *Nsegments*, that gives the ratio of the number of contour segments and the area of a colony image.

Besides shape estimates, textures have also been considered. In an image, the intensity variations which define a texture are mostly related to physical variations in the scene (such as pebbles on the ground). It is very difficult to model these variations and no precise definition of texture is present in computer vision literature [6]. For this reason, textures are usually characterized by intensity value variations in the two-dimensional space of an image. The adopted texture features of each object can be described in terms of the *gray-level co-occurrence matrix* (GLCM) [13], that allows us to capture the spatial dependence of gray-level values which contribute to the perception of texture, by showing how often different combinations of pixel brightness values occur in an image. The GLCM matrix is a square matrix $M$ of dimension $n$, where $n$ is the number of different gray levels in the image. Each element $M(i,j)$ is generated by counting the number of times a pixel in $X$ with gray value $i$ is adjacent to a pixel in $X$ having gray value $j$. Each element $M(i,j)$ is then normalized so that the sum of all elements of $M$ is equal to 1, and can thus be considered as the probability of occurrence of adjacent pixel pairs having gray level values $i$ and $j$ in the image. Among the considered texture features, the most significant are those reported in Table 1. The *Entropy* $H(X)$ measures the randomness of the gray-level distribution in the colony image $X$, while the *Contrast* $CN(X)$ is a measure of the intensity contrast between a pixel and its neighbors over $X$, assuming null value for a constant image. The *Energy* $E(X)$ yields the smallest value when all entries in $M$ are equal; it is 1 for a constant image. The *Correlation CR(X)* measures the dependency of gray levels on those of neighboring pixels. Further texture features that have been considered are the *Homogeneity*, that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal, and *Centropy* and *Bentropy*, that measure the entropy of color and grayscale colony image, respectively.

## 2.4 Classification

In machine learning, supervised classification methods aim at inferring a classification rule from a class-labeled set of examples described by a set of features. The inferred rule is then used for predicting the class of further unlabeled data. Therefore, a classifier can be considered as a mapping from a feature space to a set of classes. There exist several classifiers in literature, since there are several ways of building up such a mapping. For example, the mapping can be described by a set of induction rules, like in the tree-based classifiers, or the classifiers may be expressed as a linear separator in the original feature space, like for the perceptron or the support vector machines (SVM) classifiers. The choice of a good classifier is not a simple task, as it depends on several choices [14], including the complexity of the classification rule, the size of the training set, and the number of features.

For the classification of ESC colonies into domed and flat ones, we trained a set of binary classifiers from a set of labeled images described by the selected features (see Sect. 2.3). The classifiers have been chosen based on the criterion that the classification rule can be expressed either by a set of induction rules or by an easy interpretable mathematical function, where the importance of each feature in the classification rule is easily interpretable. This suggests which features are more relevant for labeling an image.

Specifically, we adopted three classifiers based on decision trees, that return induction rules for deciding the label of an instance: CART [2], J48 [19] and Adaboost [11] using as weak learners CART classifiers (Adaboost+CART). Among the classifiers searching for a linear separator in the feature space between instances with different labels, we used two versions of the SVM classifier: the SVM linear [24] and the SVM linear with hinge loss [18]. Finally, we adopted the Naïve Bayes classifier [15], a probabilistic algorithm which assigns a probability to each instance for each label, according to the highest label probability. All the above classifiers are available through the open source Weka software [12], that allows testing several classifiers in a very user friendly way.

## 3 Experimental Results

The proposed software framework has been tested on images of a CF assay done by plating mESCs in 96-well plates at the density of 500 cells/$cm^2$ in medium for propagation of undifferentiated cells (with Lif and serum). After cell adhesion (4 h), cells were treated with reference compounds controlling their metastability or differentiation, L-Proline [4] and Retinoic Acid [23], respectively, added at three different concentrations and allowed to grow for 4 days. Single well bright field images were acquired under a Leica MZ16FA stereo microscope.

Randomly chosen images of treated and untreated wells have been manually annotated, in order to provide the ground-truths for objective evaluation of the software components.

## 3.1 Performance Metrics

Validation of all the system modules has been performed in terms of different metrics frequently adopted in the literature [16], namely *Precision* and *Recall*

$$Recall = \frac{TP}{TP + FN}, \quad Precision = \frac{TP}{TP + FP},$$

where $TP$, $FN$, and $FP$ indicate the total number of true positives, false negatives, and false positives, respectively. *Recall*, also known as *detection rate* or *sensitivity*, gives the percentage of detected true positives as compared to the total number of true positives in the ground truth. *Precision*, also known as *positive prediction*, gives the percentage of detected true positives as compared to the total number of items detected by the method. Using the above mentioned metrics, generally a method is considered *good* if it reaches high *Recall* values, without sacrificing *Precision*. A further metric $F_1$, also known as *F-score* or *F-measure*, given by the weighted harmonic mean of *Precision* and *Recall*

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision},$$

provides a single measure that can be used to "rank" different methods.

For validating the classification accuracy, we adopted also the *False Positive Rate* (*FPR*)

$$FPR = \frac{FP}{FP + TN},$$

where $TN$ indicates the total number of true negatives.

## 3.2 Evaluation of the Segmentation Step

For the segmentation step, Recall, Precision, and $F_1$ metrics have been adopted both as pixel-based and as object-based measures, depending on wether *true positives* (resp. *false negatives*) are intended as true positive (resp. false negative) pixels or objects (colonies), respectively. While the pixel-based measures provide hints on the fine-grain segmentation accuracy, the object-based measures provide hints on the colony counting ability of the proposed modules. Average segmentation accuracy of

**Table 2** Average segmentation accuracy of the two segmentation approaches

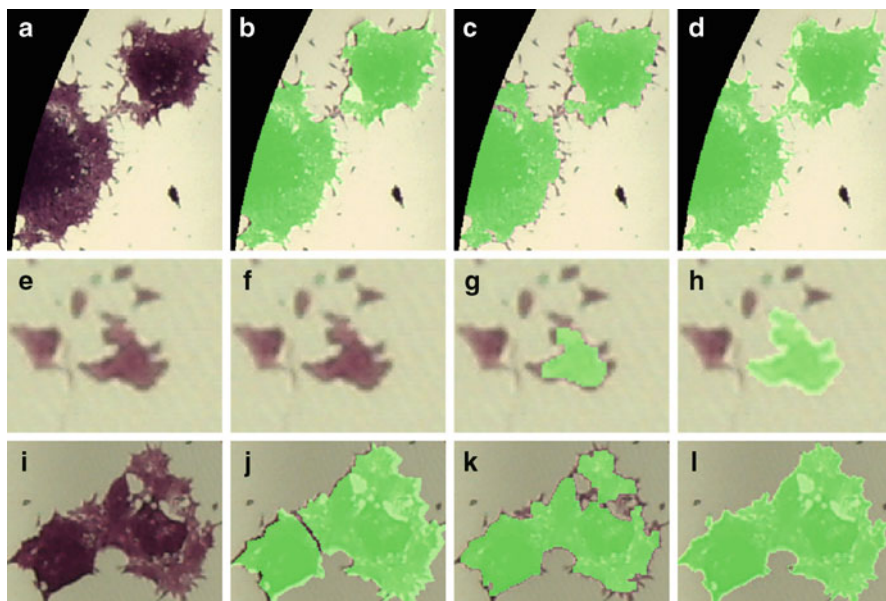| | Approach1 | | Approach 2 | |
|---|---|---|---|---|
| Metric | Pixel-based | Object-based | Pixel-based | Object-based |
| *Recall* | 0.7246 | 0.9194 | 0.8793 | 0.8387 |
| *Precision* | 0.9663 | 0.9344 | 0.8826 | 0.9630 |
| $F_1$ | 0.8282 | 0.9268 | 0.8809 | 0.8966 |



**Fig. 6** Evaluation of the segmentation step: details of the original image of Fig. 1 (first column), of the ground truth (second column), and of corresponding results obtained by the first and second segmentation approaches (third and fourth column, respectively). Green pixels superimposed on the original image indicate pixels included into the segmentation

the two segmentation approaches in term of the above metrics is reported in Table 2 and some details of the achieved results are reported in Fig. 6. The higher pixel-based accuracy values of the second approach (fourth column of Table 2) indicate a better ability of the variational model-based segmentation to precisely detect the colony contours (e.g., compare Figs. 6-g and 6-h). However, this ability leads also to fuse different colonies that are very close (e.g., see Fig. 6-d), resulting in lower object-based accuracy values (fifth column of Table 2).

Both the approaches report very few false positives and false negatives. Examples of false positives, e.g., segmented colonies that have no interest for the biologist, are reported in Figs. 6-g and 6-h, where the detected small colony has been excluded by the biologist in the ground truth reported in Fig. 6-f. Most of the false negatives are linked to adjacent colonies, barely distinguishable by an untrained human

eye, that are segmented as a single colony. An example is given in the third row of Fig. 6, where the adjacent colonies of the original image (Fig. 6-i) have been manually segmented in the ground truth as two separate colonies (Fig. 6-j) based only on the biologist experience. The two approaches, instead, both provide a single segmented colony (Figs. 6-k and 6-l). This analysis of false positives and false negatives suggests that features other than contrast or edges, able to better describe the discrimination ability of the trained eye, could help avoiding the few segmentation errors reported. Hints could be provided by the subsequent steps of feature computation and feature-based classification.

Although the segmentation results are quite accurate for both the proposed approaches, a suitable combination of them, together with a feedback from the subsequent steps of feature computation and classification, could help further improving segmentation accuracy.

## 3.3 First Classification Results

Each of the selected classifiers (see Sect. 2.4) has been trained using the selected features (see Sect. 2.3) of the colonies included into the constructed ground truths. The statistical validation of results has been obtained using a 10-fold cross validation. This procedure consists in splitting the training set in ten parts; at each step, one of those parts is used for testing and the remaining nine are used for training. Therefore, the performance metrics have been evaluated as the average among the one hundred repetitions of the ten-fold cross validation. Table 3 reports the classification performance values for each classifier, suggesting that all the six classifiers are very accurate in predicting the label of an instance. In the case of SVM with hinge loss, the number of false positives is reduced up to 1%. This is exemplified in Fig. 7, where we compare the manual classification provided by the biologist with the automatic classification computed using the SVM with hinge loss for the well of Fig. 1. Table 4 reports the weights (one for each feature) of the normalized equation of the hyperplane that separates domed vs. flat colonies in the feature space. Positive weights are associated to domed colonies. The higher the absolute value of the weights, the more the feature is relevant for discriminating

**Table 3** 10-fold cross validation of classification results

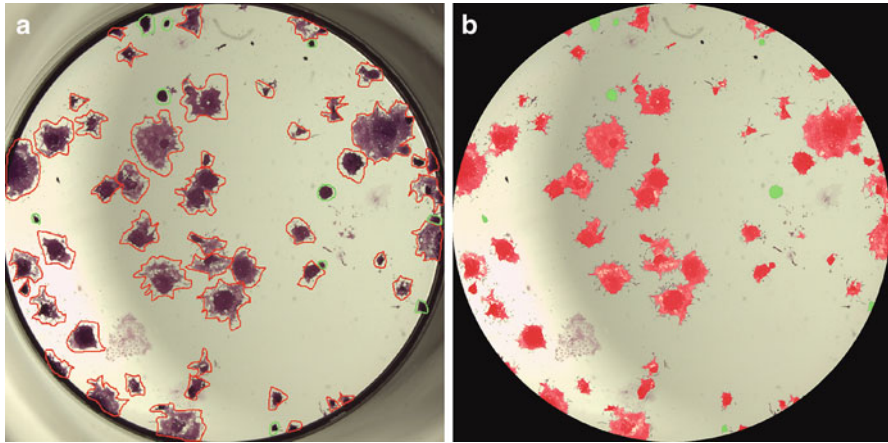| Classifier | FPR | Recall | Precision | F-measure |
|---|---|---|---|---|
| CART | 0.040 | 0.96585 | 0.96593 | 0.96587 |
| J48 | 0.032 | 0.96585 | 0.96635 | 0.96589 |
| Adaboost+CART | 0.024 | 0.97561 | 0.97563 | 0.97560 |
| SVM linear | 0.038 | 0.97073 | 0.97097 | 0.97076 |
| SVM linear with hinge loss | **0.010** | **0.98537** | **0.98540** | **0.98536** |
| Naïve Bayes | 0.024 | 0.97561 | 0.97568 | 0.97562 |

**Fig. 7** Example of classification results: (**a**) manual classification; (**b**) automatic classification using SVM with hinge loss

**Table 4** SVM with hinge loss: shape (S) and texture (T) features ordered by the absolute value of (normalised) weights

| Feature | Weight | Feature | Weight | Feature | Weight |
|---|---|---|---|---|---|
| (S) Solidity | 8.3231 | (S) Eccentricity | 3.4247 | (T) Contrast | 1.8952 |
| (S) Compactness | 7.5965 | (T) Color | −2.7622 | (S) Perimeter | −1.8120 |
| (T) Correlation | −6.4780 | (S) Min axis length | −2.6790 | (T) Homogeneity | 1.6923 |
| (S) Nsegments | 5.9170 | (T) Centropy | −2.4002 | (S) Area | −0.7793 |
| (T) Energy | 5.1318 | (S) Max axis length | −2.1554 | (T) Bentropy | 0.3709 |

the classes. Also the other classifiers[1] confirmed that, as expected, *Solidity* and *Compactness* shape features are the most discriminant features for our binary classification. Among texture features, *Correlation* is the most relevant.

## 4   Concluding Remarks

In this paper we present a novel software framework for the segmentation and classification of microscopic images obtained by CF assays of mESCs and suitable for high-throughput applications. Our main aims were to use general-purpose software components for image processing and machine learning, eventually developed in different projects, to support biological experiments within a multidisciplinary context. First results show that our approach is promising to be used as blind tool

---

[1]Detailed results may be supplied on demand.

to support reliable analysis of molecular high-throughput screening. Future works will include comparisons in terms of efficiency, reliability and flexibility with some existing application-specific software tools and the development of a user-friendly software interface for biologists without deep knowledge in computer programming.

# References

1. Bar, L., et al.: Mumford and Shah model and its applications to image segmentation and image restoration. Handbook of Mathematical Methods in Imaging, vol. I, pp. 1095–1157. Springer, New York (2011)
2. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Chapman & Hall, New York (1984)
3. Carpenter, A., Jones, T., Lamprecht, M., Clarke, C., Kang, I., Friman, O., Guertin, D., Chang, J., Lindquist, R., Moffat, J., Golland, P., Sabatini, D.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. **7**(R100) (2006)
4. Casalino, L., Comes, S., Lambazzi, G., De Stefano, B., Filosa, S., De Falco, S., De Cesare, D., Minchiotti, G., Patriarca, E.: Control of embryonic stem cell metastability by l-proline catabolism. J. Mol. Cell Biol. **3**(2), 108– (2011)
5. Celebi, M., Kingravi, H., Uddin, B., Iyatomi, H., Aslandogan, Y., Stoecker, W., Moss, R.: A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. **31**(6), 362–373 (2007)
6. Chen CH Pau LF, W.P.: The Handbook of Pattern Recognition and Computer Vision (2nd Edition). World Scientific Publishing Co, Singapore (1998)
7. Cozza, V., Guarracino, M.R., Maddalena, L., Baroni, A.: Dynamic clustering detection through multi-valued descriptors of dermoscopic images. Stat. Med. **30**, 2536–2550 (2011)
8. D'Ambra, P., Filippone, S.: A parallel generalized relaxation method for high-performance image segmentation on gpus. J. of Comput. Appl. Math. 293, 34–44 (2016)
9. D'Ambra, P., Tartaglione, G.: Solution of Ambrosio-Tortorelli model for image segmentation by generalized relaxation method. Commun. Nonlinear Sci. Numer. Simul. **20**, 819–831 (2015)
10. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM **15**(1), 11–15 (1972)
11. Freund, Y., Schapire, R.E.: A short introduction to boosting. In: In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 1401–1406. Morgan Kaufmann (1999)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
13. Haralick, R., Shanmugam, K.: Computer classification of reservoir sandstones. IEEE Trans. Geosci. Electron. **11**, 171–177 (1973)
14. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2 edn. Springer, New York (2013)
15. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345. Morgan Kaufmann, San Mateo (1995)
16. Maddalena, L., Petrosino, A.: The 3dSOBS+ algorithm for moving object detection. Comp. Vision Image Underst. **122**(0), 65–73 (2014)

17. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man and Cybern. **9**(1), 62–66 (1979)
18. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B, Burges, C, Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1998)
19. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
20. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al.: Fiji: an open-source platform for biological-image analysis. Nat. Methods **9**(7), 676–682 (2012)
21. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, Inc, New York (1983)
22. Shariff, A., Kangas, J., Coelho, L., Quinn, S., Murphy, R.: Automated image analysis for high content screening and analysis. J. Biomol. Screening **15**, 726–734 (2010)
23. Tighe, A., Gudas, L.: Retinoic acid inhibits leukemia inhibitory factor signaling pathways in mouse embryonic stem cells. J Cell Physiol. **198**(2), 223–229 (2004)
24. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
25. Zhou, X., Wong, S.T.: Informatics challenges of high-throughput microscopy. IEEE Signal Process. Mag. **23**, 63–72 (2006)