Biological cells classification using bio-inspired descriptor in a boosting k-NN framework

Wafa Bel Haj Ali, Paolo Piro, Dario Giampaglia, Thierry Pourcher, Michel Barlaud

Biological cells classification using Bio-Inspired descriptor in a boosting k-NN framework

Wafa Bel haj ali[1], Paolo Piro[2], Dario Giampaglia[1], Thierry Pourcher[3], Michel Barlaud[1]
[1]I3S - CNRS - U. Nice-Sophia Antipolis, France
[2]Italian Institute of Technology (IIT) Genova, Italy
[3]Tiro - CEA - U. Nice-Sophia Antipolis, France

Abstract

High-content imaging is an emerging technology for the analysis and quantification of biological phenomena. Thus, classifying a huge number of cells or quantifying markers from large sets of images by experts is a very time-consuming and poorly reproducible task. In order to overcome such limitations, we propose a supervised method for automatic cell classification. Our approach consists of two steps: the first one is an indexing stage based on specific bio-inspired features relying on the distribution of contrast information on segmented cells. The second one is a supervised learning stage that selects the prototypical samples best representing the cell categories. These prototypes are used in a leveraged k-NN framework to predict the class of unlabeled cells. In this paper we have tested our new learning algorithm on cellular images acquired for the analysis of pathologies. In order to evaluate the automatic classification performances, we tested our algorithm on the HEp-2 Cells dataset of (Foggia et al, CBMS 2010). Results are very promising, showing classification precision larger than 96% on average, thus suggesting our method as a valuable decision-support tool in such cellular imaging applications.

1. Introduction

Pathologists establish their diagnostics by studying tissue sections, blood samples or punctures. In general, samples are stained with various dyes to visualize cell cytoplasm and nucleus. In addition, immunohistochemistry is used to study specific protein expression. Using these approaches, pathologists observe tissue damage or cell dysfunction like for example, inflammation, neoplasia or necrosis. Abnormal nuclei allow determining cancer grades. Pathologists recognize aberrant shapes of whole cells, organelles, nuclei or staining allowing the classification of the cells. Quantification is based on visual counting. Such analysis by one (or several) experimenter is time-consuming and above all poorly reproducible. Furthermore, visual counting is generally performed on a small portion of the sample. A Computer Aided Diagnosis (CAD) system will allow reliable quantification and therefore be a precious tool in diagnostics. CAD will permit repetitive quantification on larger parts of tissue or on many cell punctures and, then, quantitative studies. In autoimmune diseases, targets of autoantibodies are characterized by indirect Immunofluorescence (IIF) on human cultured cells. Then, stained compartments of cells are identified by experts. A CAD of this analysis should provide faster and more reliable IIF. We developed a new classification method for the analysis of the staining morphology of thousands of cells. In this work, this automatic classification was used on a dataset of IIF-stained cells.

Our cell classification method consists of two steps: the first one is an indexing process based on specific bio-inspired features using contrast information distributions on cell sub-regions. The second is a supervised learning process to select prototypical samples (that best represent the cell categories) which are used in a leveraged k-NN framework to predict the class of unlabeled cells. Such classification method has many applications in cell imaging in the areas of research in basic biology and medicine but also in clinical histology.

2. Classification method

Our classification process needs two major steps as shown in Fig. 1: first we compute bio-inspired descriptors, extracting contrast-based features in the segmented cells. These descriptors are then used in a supervised learning framework where the most relevant prototypical samples are used to predict the class of unlabeled cells. We split this section in two parts: the first describes our feature extraction approach, whereas the latter is focused on our

prototype-based learning algorithm.

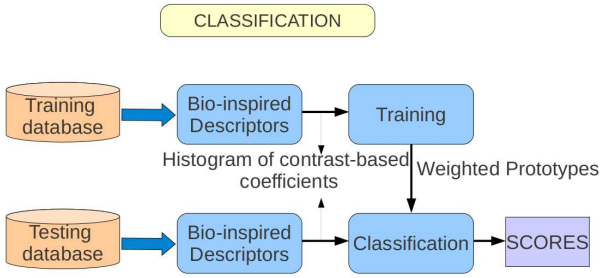

Figure 1. Block scheme of our cell classification method.

2.1. Region based bio-inspired descriptor

For better understanding the image content, it can be useful to get inspiration from the way our visual system operates to analyze the scene. The first transformation undergone by a visual input is performed by the retina.

In fact, ganglion cells, that are the final output of the retina, are first simulated by the local changes of the *illumination*. This information is captured by their receptive fields and transformed to *luminance contrast* intensities. Those receptive fields are like center-surround models (see Fig. 2). They react to the illumination of either the center or the surround of the ganglion cells and are disabled when illuminating the other one. Such behavior, similar to an edge detector, is modeled by a centered two-dimensional *Difference of Gaussians* (1).

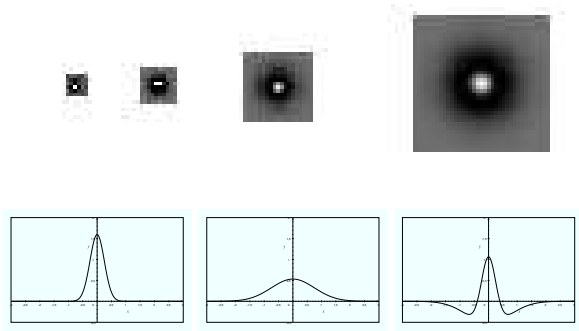

Figure 2. Top, receptive fields in the retina modeled by DoGs for 4 scales. Bellow, the model of the response of those retinal cells.

$$DoG_\sigma(x,y) = G_\sigma(x,y) - G_{\alpha\cdot\sigma}(x,y) \qquad (1)$$

Moreover, ganglion cells react to the luminance in different scales, thus adding multiscale aspect and allowing us to use DoG filters in a scale space (Fig. 2).

The basic idea is to compute features inspired from the visual system model and specially from the main characteristics of the retina processing. Such was the case in [1], where we represented the image using features based on *contrast* information on square blocs.

Such descriptor is well adapted in the case of our cells images since the most discriminative visual feature between categories is the *luminance contrast* in subcellular regions. Thus, we define cell descriptors based on the *local contrast* in the cell, that we call Bio-Inspired Features, BIF. The *local contrast* is obtained by a filtering with *Differences of Gaussians* (DoGs) centered at the origin. So that the contrast C_{Im} for each position (x, y) and a given scale s in the image Im is as follows:

$$C_{Im}(x,y,s) = \sum_i \sum_j (Im(i+x, j+y) \cdot DoG_{\sigma(s)}(i,j)) \qquad (2)$$

We use the DoG described by [2] where the larger Gaussian has three times the standard deviation of the smaller one. After computing these contrast coefficients in (2), we apply a non-linear bounded transfer function, named neuron *firing rates*, used in [10]. This function is written as:

$$R(C) = G \cdot C/(1 + Ref \cdot G \cdot C), \qquad (3)$$

where G is named the contrast gain and Ref is known as the refractory period, a time interval during which a neuron cell *reacts*. The values of those two parameters proposed in [10] to best approximate the retinal system are $G = 2000\,Hz \cdot contrast^{-1}$ and $Ref = 0.005\,s$.

Firing rate coefficients $R(C)$ are encoded on an already segmented cell region. Then, they are quantified into normalized $\mathcal{L}1$ histograms of n-bins for each scale and finally concatenated. Thus our global descriptor's dimension is a multiple of n.

Note that state of the art classical methods such as SIFT descriptors encode gradient directions on square blocks [4]

2.2. Prototype-based learning

We consider the multi-class problem of automatic cell classification as multiple binary classification problems in the common one-versus-all learning framework [8]. Thus, for each class c, a query image is given a positive (negative) membership with a certain confidence (classification score). Then the label with the maximum score is assigned to the query.

We suppose given a set $\mathcal{S}$ of m annotated images. Each image is a training *example* $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ is the image

feature vector and $\boldsymbol{y} = \{-1, 1\}^C$ the *class vector* that specifies the category membership of the image. In particular, the sign of component y_c gives the positive/negative membership of the example to class c ($c = 1, 2, ..., C$), such that y_c is negative iff the observation does not belong to class c, positive otherwise.

In this paper, we propose to generalize the classic k-NN rule to the following *leveraged* multiclass classifier $\boldsymbol{h}^\ell = \{h_c^\ell\}$:

$$h_c^\ell(\boldsymbol{x}_q) = \sum_{j=1}^{T} \alpha_{jc} K(\boldsymbol{x}_q, \boldsymbol{x}_j) y_{jc} \ , \tag{4}$$

where h_c^ℓ is the classification score for class c, $\boldsymbol{x}_q$ denotes the query image, α_{jc} the *leveraging coefficients*, which provide a *weighted* voting rule instead of uniform voting, and $K(\cdot, \cdot)$ is the k-NN indicator function:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1 \ , & \boldsymbol{x}_j \in \mathrm{NN}_k(\boldsymbol{x}_i) \\ 0 \ , & \text{otherwise} \end{cases} \ , \tag{5}$$

with $\mathrm{NN}_k(\boldsymbol{x}_i)$ denoting the set of the k-nearest neighbors of $\boldsymbol{x}_i$.

Training our classifier essentially consists in selecting the most relevant subset of training data, *i.e.*, the so-called *prototypes*, whose cardinality T is generally much smaller than the original number m of annotated instances. The prototypes are selected by first fitting the coefficients α_j, and then removing the examples with the smallest α_j, which are less relevant as prototypes.

In order to fit our leveraged classification rule (4) onto training set $\mathcal{S}$, we should try to directly minimize the multiclass surrogate[1] (exponential) risk, which is defined as the actual misclassification rate on the training data, as follows:

$$\varepsilon^{\exp}\left(h_c^\ell, \mathcal{S}\right) \doteq \frac{1}{m} \sum_{i=1}^{m} \exp\left\{-\rho(h_c^\ell, i)\right\} \ , \tag{6}$$

where:

$$\rho(h_c^\ell, i) = y_{ic} h_c^\ell(\boldsymbol{x}_i) \tag{7}$$

is the multiclass *edge* of classifier h_c^ℓ on training example $\boldsymbol{x}_i$. This edge measures the "goodness of fit" of the classifier on example $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for class c, thus being positive iff the prediction agrees with the example's annotation.

In order to solve this optimization, we propose a boosting-like procedure, *i.e.*, an iterative strategy where the classification rule is updated by adding a new prototype $(\boldsymbol{x}_j, \boldsymbol{y}_j)$ (weak classifier) at each step t ($t = 1, 2, \ldots, T$), thus updating the strong classifier (4) as follows:

$$h_c^{(t)}(\boldsymbol{x}_i) = h_c^{(t-1)}(\boldsymbol{x}_i) + \delta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) y_{jc} \ . \tag{8}$$

[1] We call *surrogate* a function that upperbounds the risk functional we should minimize, and thus can be used as a primer for its minimization.

(j is the index of the prototype chosen at iteration t.) Using (8) into (7), and then plugging it into (6), turns the problem of minimizing (6) to that of finding δ_j with the following objective:

$$\arg\min_{\delta_j} \sum_{i=1}^{m} w_i \cdot \exp\left\{-\delta_j \mathrm{r}_{ij}\right\} . \tag{9}$$

In (9), we have defined r_{ij} as a pairwise term only depending on training data:

$$\mathrm{r}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) y_{ic} y_{jc} \ . \tag{10}$$

and w_i as the weighting factor, depending on the past weak classifiers:

$$w_i = \exp\left\{-y_{ic} h_c^{(t-1)}(\boldsymbol{x}_i)\right\} , \tag{11}$$

Finally, taking the derivative of (9), the global minimization of surrogate risk (6) gives the following expression of δ_j :

$$\delta_j = \frac{1}{2} \log \frac{\gamma \cdot \sum_{i: r_{ij}^c > 0} w_i}{\sum_{i: r_{ij}^c < 0} w_i} \ , \tag{12}$$

where γ is a coefficient that compensates for the imbalance between positive and negative examples.

We provided theoretical details and properties of our boosting algorithm in [7], as well as an extension of UNN to inherent multiclass classification in [6].

We also tried a "soft" version of the UNN classification rule, called UNN_s, which considers a logistic estimator for a Bernoulli prior that vanishes with the rank of the neighbors, thus decreasing the importance of farther neighbors:

$$\hat{p}(j) = \beta_j = \frac{1}{1 + \exp\left(\lambda(j-1)\right)} \ . \tag{13}$$

This amounts to redefining (4) as follows:

$$h_c^\ell(\boldsymbol{x}_q) = \sum_{j=1}^{T} \alpha_{jc} \beta_j K(\boldsymbol{x}_q, \boldsymbol{x}_j) y_{jc} \ . \tag{14}$$

(Notice that k-NN indexed by j are supposed to be sorted from closer to farther.)

3. Experiments

We evaluated our classification approach on the HEp-2 Cells dataset [3] provided by University of Salerno and Campus Bio-Medico of Roma[2]. This database contains 721 images divided into *six* categories as shown in Fig. 3. Cells

[2] Data available at: `http://mivia.unisa.it/hep2contest/index.shtml`

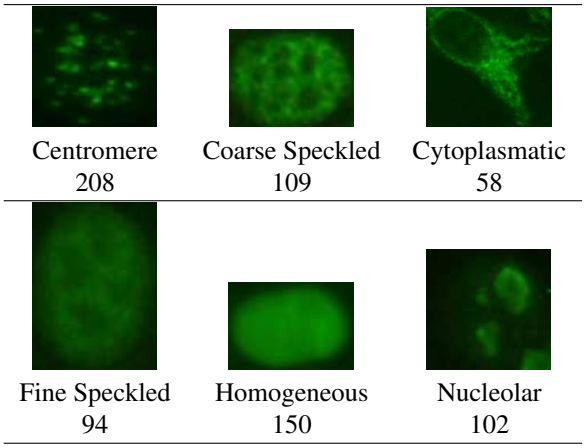

Figure 3. Sample images and the number of elements for each category in the dataset.

are already segmented and both hole images and their corresponding masks are provided in the dataset.

In a first step, we extract *Bio-Inspired* features for each segmented cell according to the cell mask. Some parameters, such as scales and dimension of the descriptor, should be tuned. For this purpose, we carried out an analysis of the classification precision as a function of scales and descriptor dimension, as reported in Tab. 1. Our global features are the

N of bins	*32*	*64*	*128*	*256*	*512*
1 scale	61.50	71.59	79.02	83.83	86.32
2 scales	87.51	91.01	91.84	92.90	92.81
3 scales	90.91	94.06	95.32	95.51	95.44
4 scales	93.69	95.53	**96.04**	**96.08**	96.03
5 scales	94.33	95.56	95.94	95.80	95.69

Table 1. UNN **classification rate as a function of the scale and the bin's number of the descriptor.**

concatenation of histograms of n-bins for each scale. The dimension is then equal to the number of scales multiplied by n. The cross validation experiment in Tab. 1, obtained with UNN algorithm, shows that using 4 scales with a bins number equal to 256 or even 128 gives best performances. Thus, next evaluations are performed using the global dimension equal to 1024 for descriptors.

We compare our descriptor to the state of the art SIFT [4]. SIFT features encode gradient directions on small square blocks of subcellular regions. However gradient directions are not relevant features for such biologic cells, thus using SIFT leads to poor classification rates compared to those obtained using our approach. Next, we use Bag-of-Words [9] (BoW) with a dimension equal to 1024 build with dense SIFT of [11].

We evaluate performances using cross validation on 100 folds. For each fold we randomly choose 50% of the images for training, while testing on the remaining ones. We use the *TP rate* (True positive rate) and the *AUC* (Area under roc curve) as measures of classification precision. Both of them are computed by averaging over tests on the 100 random folds.

We compared performances of UNN with those of standard k-NN and SVM, using Bio-Inspired descriptors and BoW with SIFT ones. The *TP rates* and *AUC* per category and the average *TP rate* and *AUC* (last columns) are reported respectively in Tab. 2 and Tab. 3. These results display the high discriminative ability of the proposed Bio-Inspired descriptor, which allows for classification precision generally larger than 90%, up to almost 100% (on the "Coarse Speckled" and "Cytoplasmatic" classes). Contrast based descriptors are more relevant features than gradient based SIFT for biological cells classification. Our bioinspired descriptor outperforms classical SIFT state of the art descriptor for all classification methods. Furthermore, our UNN classification method improves the classic k-NN, most significantly on the "Fine Speckled" and "Homogeneous" classes, with improvement larger than 2%. At the same time, our learning method achieves performances always comparable with those of state-of-the-art SVM. For instance, notice the improvement of UNN over SVM on the "Coarse Speckled" class (2.5% gap on *TP* and 1.5% on *AUC*) and the "Fine Speckled" one (2% gap on *TP* and 4% on *AUC*), while SVM is the best performing method on the "Homogeneous" and "Nucleolar" classes. For further

	Centromere	Coarse Sp.	Cytoplasm.	Fine Sp.	Homogen.	Nucleolar
Centromere	96.33	0.50	0.91	0.82	0.50	0.91
Coarse Sp.	0.18	98.03	0.68	0.85	0.22	0.01
Cytoplasm.	0	0.20	99.55	0.10	0.10	0.03
Fine Sp.	0.78	0.42	0.12	95.36	3.29	0
Homogen.	0.65	2.50	0.60	5.32	90.84	0.08
Nucleolar	1.37	2.15	0	0	0.05	96.41

Table 4. Confusion table for UNN**: the average classification rate (or mAP) is the mean of the diagonal of this matrix.**

details on the classification precision per category we give as example the confusion matrix for UNN classification in Tab. 4.

Besides comparing very favorably with state-of-the-art approaches, our UNN method enables much faster classification. Fig. 4, shows typical processing time for UNN and

	Centromere	Coarse Speckled	Cytoplasmatic	Fine Speckled	Homogeneous	Nucleolar	TP rate
k-NN + BIF	94.88	98.77	98.44	86.65	83.88	94.11	92.79
UNN + BIF	96.33	98.03	99.55	95.36	90.84	96.41	96.08
SVM + BIF	96.38	95.68	99.86	93.36	93.90	97.92	96.18
k-NN + SIFT	86.09	45.35	100	41.61	94.78	71.49	73.22
UNN + SIFT	86.13	69.22	99.89	63.59	91.30	85.62	82.63
SVM + SIFT	88.22	70.87	98.41	62.36	87.52	90.47	82.97

Table 2. Performances of k-NN, UNN and SVM with BIF and SIFT descriptors in term of *TP rate* (true positif rate) for each of the six classes.

	Centromere	Coarse Speckled	Cytoplasmatic	Fine Speckled	Homogeneous	Nucleolar	AUC
k-NN + BIF	94.58	90.26	100	51.60	84.58	84.49	84.25
UNN + BIF	95.12	96.78	99.03	95.78	93.91	94.69	95.89
SVM + BIF	95.96	95.21	98.27	91.67	97.24	98.68	96.17
k-NN + SIFT	95.06	90.35	99.70	57.61	81.97	89.01	85.61
UNN + SIFT	92.46	87.23	98.21	66.11	90.00	91.09	87.52
SVM + SIFT	92.18	79.06	92.56	60.00	91.76	92.57	84.69

Table 3. Performances of k-NN, UNN and SVM with BIF and SIFT descriptors in term of *AUC* (area under RoC curve) for each of the six classes.

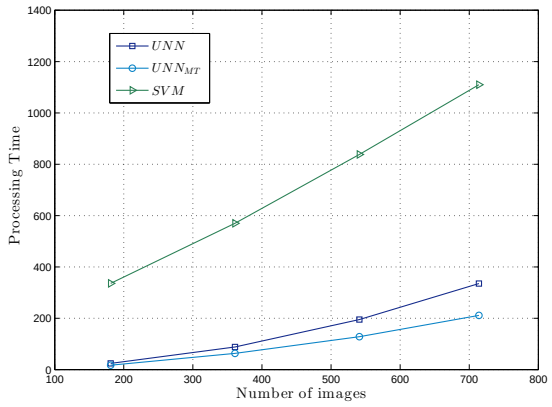

Figure 4. Processing time of the training step for both UNN, SVM and multi-thread version of UNN.

SVM and UNN achieves speedups of roughly 3 to 5 over SVM. UNN benefits from straightforward multi-thread implementation (UNN_{MT}) in addition to the fast and efficient tool that we used for the k-NN search algorithm, provided in the Yael toolbox[3]. This makes the processing furthermore faster. Therefore our Bio-Inspired UNN approach provides the best mAP/Time trade-off.

[3]Source code available at: `https://gforge.inria.fr/projects/yael`

4. Conclusion

In this paper, we have presented a novel algorithm for automatic supervised classification of cellular images. First of all, our method relies on extracting highly discriminative descriptors based on Bio-Inspired histograms of Difference-of-Gaussians (DoG) coefficients on cellular regions. Then, we propose a supervised classification algorithm, called UNN, for learning the most relevant prototypical samples that are to be used for predicting the class of unlabeled cellular images according to a leveraged k-NN rule. We evaluated UNN performances on the HEp-2 Cells dataset (manually segmented and annotated). Although being the early results of our methodology for such a challenging application, performances are really satisfactory (average global precision of 96%).

References

[1] W. Bel haj ali, P. Piro, L. Crescence, D. Giampaglia, O. Ferhat, J. Darcourt, T. Pourcher, and M. Barlaud. A bio-inspired learning and classification method for subcellular localization of a plasma membrane protein. In *International Conference on Computer Vision Theory and Applications (VISAPP 2012)*, 2012.

[2] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, 1994.

[3] P. Foggia, G. Percannella, P. Soda, and M. Vento. Early experiences in mitotic cells recognition on hep-2 slides. In *CBMS'10*, pages 38–43, 2010.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001. 10.1023/A:1011139631724.

[6] P. Piro, R. Nock, F. Nielsen, and M. Barlaud. Multi-Class Leveraged k-NN for Image Classification. In *Proceedings of the Asian Conference on Computer Vision (ACCV 2010)*, 2010.

[7] P. Piro, R. Nock, F. Nielsen, and M. Barlaud. Leveraging k-nn for generic classification boosting. *Neurocomputing*, 80(0):3–9, March 2012.

[8] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

[9] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 127–144. Springer Berlin / Heidelberg, 2006. 10.1007/11957959_7.

[10] R. Van Rullen and S. J. Thorpe. Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Comput*, 13(6):1255–1283, June 2001.

[11] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.