BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 94–101
doi:10.1093/bioinformatics/btm530

Systems biology

Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy

Meng Wang[1], Xiaobo Zhou[1,2,*], Fuhai Li[1], Jeremy Huckins[3], Randall W. King[3] and Stephen T.C. Wong[1,2]

[1]Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair, Harvard Medical School, 3rd floor, 1249 Boylston, Boston, MA 02215, [2]Functional and Molecular Imaging Center, Department of Radiology, Brigham and Women's Hospital, One Brigham Circle, 1620 Tremont Street, Boston, MA 02121 and [3]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

Received on April 8, 2007; revised on September 12, 2007; accepted on October 16, 2007

Advance Access publication November 7, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Automated identification of cell cycle phases captured via fluorescent microscopy is very important for understanding cell cycle and for drug discovery. In this article, we propose a novel cell detection method that utilizes both the intensity and shape information of the cell for better segmentation quality. In contrast to conventional off-line learning algorithms, an Online Support Vector Classifier (OSVC) is thus proposed, which removes support vectors from the old model and assigns new training examples weighted according to their importance to accommodate the ever-changing experimental conditions.

Results: We image three cell lines using fluorescent microscopy under different experiment conditions, including one treated with taxol. Then, we segment and classify the cell types into interphase, prophase, metaphase and anaphase. Experimental results show the effectiveness of the proposed system in image segmentation and cell phase identification.

Availability: The software and test datasets are available from the authors.

Contact: zhou@crystal.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The knowledge of cell cycle progress, e.g. interphase, prophase, metaphase and anaphase, is important in understanding various diseases, notably cancer (Yan *et al*., 2006; Zhou and Wong, 2006). Changes in the cell cycle before and after drug use are useful for effective drug discovery research (Anderson *et al*., 2003; Baguley and Marshall, 2004; Dixon *et al*., 2002). Cell cycle progress can be identified by measuring changes in the nucleus as a function of time. Automated fluorescence microscopy imaging provides an important method to study nuclei dynamically and thus becomes an important quantitative technique in the fields of cell and systems biology (Chen *et al*., 2006; Harder *et al*., 2006; Wang *et al*., 2007; Yan *et al*., 2006; Zhou and Wong, 2006). Figure 1 (Fig. 4 in Wang *et al*., 2007) provides an example of cell mitosis process. Murphy *et al*. (Boland and Murphy, 2001; Boland *et al*., 1998; Chen and Murphy, 2006; Huang *et al*., 2003; Murphy *et al*., 2003) have proposed different feature extraction, feature reduction and classification algorithms for a similar problem of classification of subcellular location patterns in fluorescence microscope images. Although some methods have been proposed for cell cycle phase recognition (Chen *et al*., 2006; Gallardo *et al*., 2004), the automatic stratification of different phases of cells is still an unresolved issue in cell biology studies using fluorescence microscopy (Yan *et al*., 2006; Zhou and Wong, 2006). Along these lines, an automated system also based on the Support Vector Classifier (Charles *et al*., 2007; Harder *et al*., 2006) was proposed recently for classifying seven mitotic phases system. However, its classifier is updated in a 'batch' mode in contrast to our classifier's 'online' mode. A context-based model (Wang *et al*., 2007) has been proposed to deal with this issue. However, its performance is highly dependent on the robustness of the tracking algorithm. When the cells are over-populated, the tracking algorithm becomes less reliable and the context information becomes less informative. Thus, it is important to develop a classification algorithm that is independent of the context information.

In this article, we present new segmentation and online learning algorithms to acquire and analyze cell cycle behaviors of a population of cells whose image is generated by microscopy. Here, online refers to the fact that the classifier will be continually updated by the misclassified samples to accommodate the ever-changing experimental conditions. Considering cells may cluster or overlap with each other, a novel cell detection algorithm is thus proposed. Figure 2 shows the detailed flowchart of the proposed segmentation method. First, cell shape information is obtained with a binarization process (Lindblad *et al*., 2004; Wahlby *et al*., 2002). Second, both intensity and shape information is used for local maxima generation. Finally, the local maxima (centers of cells) are detected inside the gradient vector field (the pixels will

*To whom correspondence should be addressed.

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

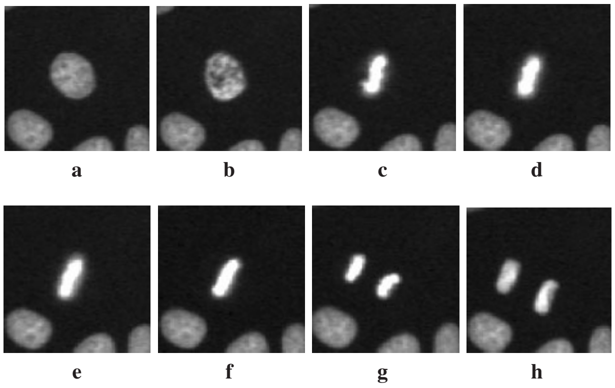

Fig. 1. Changes in the appearance of a nucleus during cell mitosis. From (**a**) to (**h**) consecutive image subframes form a sequence showing nuclei size and shape changes during cell mitosis.

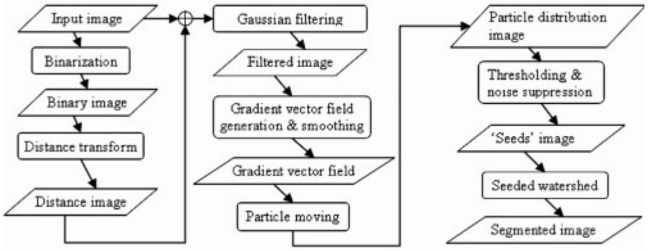

Fig. 2. The flowchart of the proposed procedure.

ultimately converge at these local maxima), and the detected cells are then segmented via a seeded watershed algorithm (Lin *et al*., 2003). After segmentation, the favorite matching plus local tree matching approach is used to track the dynamic behaviors of cell nuclei (Yan *et al*., 2006).

After obtaining the segmented nuclei, each cell is represented with a feature vector. Each feature vector contains 211 features: 10 general image features about shape, size and intensity (Chen *et al*., 2006); 14 Haralick co-occurrence textural features (Haralick *et al*., 1973); 47 Zernike moment features (Boland and Murphy, 2001; Gallardo *et al*., 2004); 85 features generated by Gabor transformation (Manjunath and Ma, 1996; Zhou and Wong, 2006) and 54 shape features. After feature selection, 58 features are kept as the new feature vector for phase identification.

While the predominant approach focuses on the development of off-line classifiers to improve the classification performance, we observe that the online adaptivity is necessary to cope with new problems facing high-throughput cellular imaging, including the ever-changing experimental conditions, the drifting feature values after treatment with anti-mitotic drugs such as taxol. Meanwhile, nuclear morphology also requires the algorithm that has the characteristic of online adaptivity (Debes *et al*., 2005; Stern *et al*., 2005). The Perceptron (Rosenblatt, 1958) was known as the first simple and efficient online learning algorithm. After that, another online kernel classifier (Freund and Schapire, 1999) was proposed based on the same principle. Support Vector Machines (SVMs) (Cortes and Vapnik, 1995; Guyon *et al*., 1993; Vapnik, 1998; Vapnik and Lerne, 1963) are the successful application of the kernel idea to large margin classifiers. Conventionally, Support Vector Classification (SVC) has been used in the batch setting. Recently, several online algorithms have been proposed (Borders, 2005; Kivinen *et al*., 2004; Lau and Wu, 2003), which differ in the optimization and update strategy. LASVM (Borders, 2005) features the simplicity and efficiency. Shalev-Shwartz and Singer (2006) describe a new framework for the design and analysis of online learning algorithms based on the notation of duality in constrained optimization. The process of online learning can be reduced to the task of incrementally increasing the dual objective function. The image datasets for training classifiers are severely imbalanced. For example, a typical 200 frames sample of microscope images contains at least 18 000 interphase cells, while the other types of cells sum up to less than 1000. Inspired by LASVM (Borders, 2005) and the framework suggested by Shalev-Shwartz and Singer (2006), we propose an online support vector classifier to solve the problem caused by the imbalanced cell image datasets and to improve the prediction accuracy of the classes of interest.

2 METHODS AND ALGORITHMS

2.1 Cell culture

HeLa H2B-GFP cells were thawed 6 days before plating for each experiment and cultured in DMEM with 10% FBS. Cells were incubated at 37°C in 5% CO_2. All cells were plated in 8 well #1 German borosilicate sterile bottomed plates (Nalge Nunc International) 18 h before imaging at 25 000 cells per well (50 000 cells per ml). Untreated cells received medium while treated cells received 150 nm Taxol.

2.2 Image acquisition

Images were acquired on an automated epifluorescence TE2000-E Eclipse microscope (Nikon Instruments Inc., USA) with a motorized XYZ-plane stage. Light was from a mercury arc lamp with two neutral density filters. SimplePCI was used to control image acquisition. A custom designed microscope incubator set at 37°C was used to keep a constant environment while acquiring images. Representative fields were chosen and the starting *X*, *Y* and *Z* coordinates were used to seed the focusing position. The microscopy will refine the focusing position by itself at the first pass of every 10 passes to compensate for deviation of focusing position. Images were acquired using a 0.2 s exposure time, every 15 min for 50 h, giving a total of 200 images for each position, which were then exported from SimplePCI as 16 bit uncompressed TIFF files to a 7 TB network attached storage (NAS) arrays for processing with CellIQ.

2.3 Image segmentation and tracking

Image segmentation quality directly affects the tracking and cell phase recognition performance. Herein, we propose a cell segmentation method that consists of three major steps: binarization, cell detection and seeded watershed-based segmentation. Each step generates the input images for the next step. Figure 2 gives a detailed flowchart of the proposed segmentation method. Cell detection is the most important step, and generates the 'seeds' image of the seeded watershed algorithm, thereby determining the segmentation results.

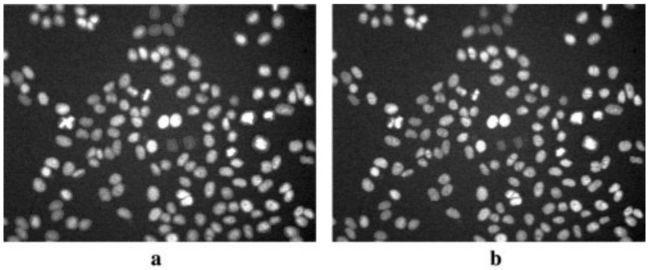

Fig. 3. Segmentation results of the proposed cell segmentation algorithm. The nuclear boundaries are shown in red. (**a**) Segmentation result with the proposed method. (**b**) Segmentation result with traditional watershed-based method (Yan *et al.*, 2006).

2.3.1 Binarization To generate the binary image for the distance transform (Breu *et al.*, 1995), we need to separate the cells from the background. It is well known the global threshold cannot generate good binary image when the intensity of background is uneven. Herein, we propose a binarization process using an adaptive thresholding method that based on the fact that there is an obvious intensity jump between the background and the object. We employ a data-driven background correction algorithm (Lindblad *et al.*, 2004; Wahlby *et al.*, 2002) to estimate the background with cubic B-spline (Supplementary Material). We classify each pixel as part of the object if the difference between its intensity and that of the estimated background image is greater than a given threshold; otherwise, we classify it as part of the background.

2.3.2 Cell detection Local intensity maxima are often used in cell detection. The distance image obtained by applying the distance transform on the binary image gives the cells' shape information. To utilize these two kinds of information, the original image is added to the distance image, which is formulated as: $I_1 = I_0 + \alpha I_{\mathrm{Dis}}$, where I_1 is the new image, I_0 the original image and I_{Dis} the distance image obtained by applying the distance transform on the binary image. The parameter α is experimentally set to 0.4. After that, the new image I_1 is filtered with Gaussian filtering, a good smoother, with SD $\sigma = 3$ (Steger, 1998). In the filtered image, the noise is suppressed and the local maxima may correspond to the cell centers. Thus, the cell detection problem can be reduced to detecting local maxima in the filtered image.

It is well known that in the gradient vector field (GVF), the gradient vectors point to the local maxima. Analogous to the electron moving inside the electron field, we put one particle on each object pixel and allow the particles to move inside the GVF. To implement this process, given one particle at a pixel, we move the particle along the gradient vector of the pixel; if the gradient vector points to another pixel, the particle will move toward another pixel, while if the gradient vector points to the pixel itself, the particle will stay still. We repeat this process for each particle until the particles stop at the local maxima (Supplementary Material**)**. To reduce the false local maxima, the gradient vector field is smoothed with the method proposed in Xu and Prince (1998), which minimizes an energy function to reach this end (Supplementary Material). Therefore, after moving the particle, the local maxima can easily be detected by thresholding the number of pixels accumulated at these points since no or very few particles accumulate at non-maxima and noises points. The centers of cells are represented by the detected local maxima.

The cells are then segmented via seeded watershed (Lin *et al.*, 2003). Figure 3a gives one segmentation result using the detected local maxima as seeds, while Figure 3b is the result using the watershed method and fragments merging (Yan *et al.*, 2006) (Supplementary Material**)**. It can be easily observed that the proposed algorithm has fewer over segmentation errors. The detailed comparison is given in Section 3.

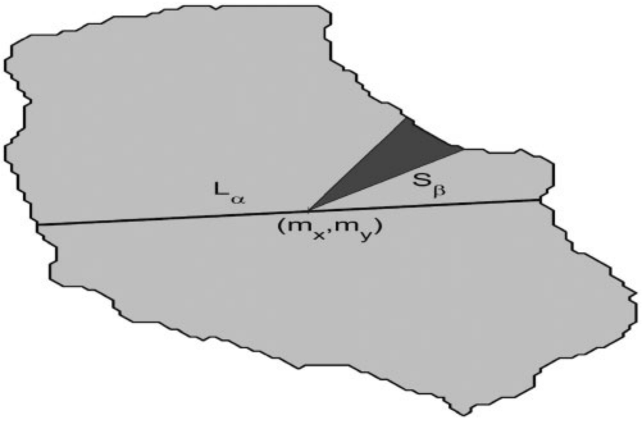

Fig. 4. The illustration of the definition of shape descriptor.

After segmentation, the favorite matching plus local tree matching approach is used to track the dynamic behaviors of the cell nuclei (Yan *et al.*, 2006). After computing the similarity scores between all possible pairs of cell nuclei in frames t and $t+1$, we search the most favorite cell in frame $t+1$ for each cell in frame t and vice versa. If two cells from frame t and $t+1$ match with each other, then we treat them as the matched pair. If all pairs of cells are treated as a graph and filtered with the favorite matching, we search all the connected sub-graphs in frames t and $t+1$ and match them by optimal tree structure searching. The cells left in these two frames can be matched by a set of heuristic rules (Yan *et al.*, 2006; Zhou and Wong, 2006) (Supplementary Material). The statistical distributions of cell's behaviors can be found in Zhou and Wong, 2006.

2.4 Feature extraction and feature selection

After obtaining the segmented nuclei, feature vectors are generated to represent the cells. Each feature vector contains 211 features. These features are composed of 10 general image features about shape, size and intensity (max intensity, min intensity, deviation of gray level, average intensity, length of long axis, length of short axis, long axis/short axis, area, perimeter) (Chen *et al.*, 2006); 14 Haralick co-occurrence textural features (Haralick *et al.*, 1973); 47 Zernike moment features (Boland and Murphy, 2001; Gallardo *et al.*, 2004); 85 features generated by Gabor transformation (Manjunath and Ma, 1996; Zhou and Wong, 2006) and 54 shape features.

We develop one category of features. One kind of feature is based on the radii drawn at a series of 36 different angles (of 10 degrees each) through the centroid normalized by the nuclear perimeter, while the other consists of the 18 areas between each pair of radii normalized by the nuclear area. They are illustrated in Figure 4 (Supplementary Material). It is worth noting that the radii are drawn at a series of 18 different angles (of 20 degrees each) in the second kind of feature. Finally, we obtain a shape descriptor with 54 elements and 4 have been used in the final feature subsets after feature selection.

To remove the irrelevant features and improve the performance of the learning system, a prediction risk-based feature selection method is employed to choose the sub-optimal feature sets (Guyon *et al.*, 2002; Li *et al.*, 2004). This method employs an embedded feature selection criterion of prediction risk, which evaluates features by calculating the change if the corresponding feature is replaced by its average value. It has several advantages. (1) The embedded feature selection model depends on learning machines. It can reach higher accuracy than the filter model, but it features lower computation complexity than the wrapper model. (2) The prediction risk criterion had been employed

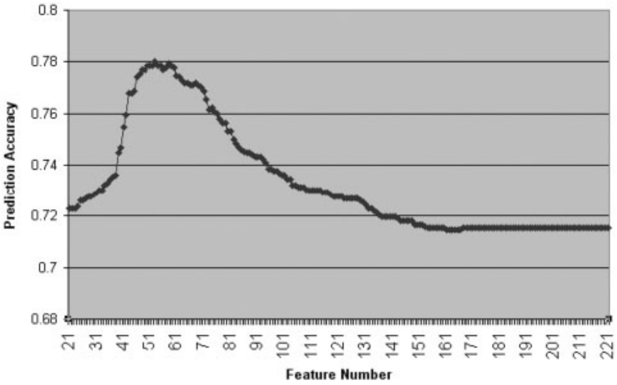

Fig. 5. The trend of the prediction accuracy while increasing the feature number.

with several different learning machines (Li *et al.*, 2004) and outperformed Optimal Brain Damage (Guyon *et al.*, 2002) when using multi-class SVMs to test on more than 10 University of California Irvine (UCI) datasets. (3) This method is easily implemented. Fifty-eight features are kept for cell phase identification, comprised of 37 Gabor features, 1 geometric feature, 14 moment features, 2 texture features and 4 shape features. The geometric feature used is ‘perimeter’. Gabor features can describe the nuclear both in the time and frequency domain. Thus so many Gabor features are kept. Figure 5 illustrates the trends of the prediction accuracy while decreasing the number of features used in the prediction algorithm.

2.5 Online support vector classifier

The basic idea of applying SVMs into interphase, prophase, metaphase and anaphase can be outlined as follows. First, map the input vectors into a feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant to the selection of the kernel function. Then, within the feature space, seek an optimized linear division; i.e. construct a hyper-plane that can separate the entire samples into two classes (this can be extended to multi-classes) with the least errors and maximal margin. The SVMs training process always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description of the theory of SVMs for pattern recognition is given in the book by Vapnik (1998).

2.5.1 Problem formulation Given a set of l samples, i.e. a series of input vectors $\mathbf{x}_i \in \mathbf{R}^d$ $(i = 1, \ldots, l)$, where x_i is the ith vector, and $\mathbf{R}^d$ is a Euclidean space with d dimensions. Suppose the output is expressed by $y_i \in \{+1, -1\}$ $(i = 1, \ldots, l)$, where the indexes -1 and $+1$ represent respectively the two classes. The primal problem of SVM (Vapnik, 1998) is given as following:

$$\begin{aligned} &\min_{\omega,b} P(\omega) = \frac{1}{2}\omega^T\omega + C\sum_{t=1}^{1}\xi_i \\ &\text{s.t. } y_i\,(w^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ &\xi_i \geq 0, i = 1, \ldots, l \end{aligned} \tag{1}$$

where C is a constant, ξ_i are slack variables, b is the bias term, ω is the weight vector and φ is a function that maps an example into the feature space.

We use the Lagrange multiplier method to solve the above optimization problem:

$$\begin{aligned} &L(\omega, \xi_i, b, \alpha_i, \beta_i) = \\ &\frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}[\alpha_i(y_i(w^T\phi(\boldsymbol{x}_i) + b) + \xi_i) + \beta_i\xi_i]. \end{aligned} \tag{2}$$

where $\alpha_i \geq 0$, $\beta_i \geq 0$, $\delta \geq 0$ are all Lagrange multipliers. Its dual problem can thus be obtained:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = &\sum_{i=1}^{l} a_i - \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j \cdot K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\text{s.t.} \quad 0 \leq a_i \leq C, i = 1, \ldots, l \\ &\sum_{i=1}^{l} a_i y_i = 0 \end{aligned} \tag{3}$$

The gradient of $D(\alpha)$ is denoted as $g = (g_1 \ldots g_n) \cdot g_k$ is calculated by:

$$g_k = \frac{\partial D(\boldsymbol{\alpha})}{\partial \alpha_k} = y_k - \sum_i \alpha_i y_i K_{i,j} = y_k - \hat{y}(x_k) + b \tag{4}$$

where $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function, and $\hat{y}$ is the predicted value.

In this article, the RBF kernel $k(x,x') = \exp(-||x - x'||^2/\gamma)$ is used. The decision function is given by:

$$\tilde{f}(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{l} y_i\alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{5}$$

Various numerical algorithms have been developed to solve the SVM QP problem (Platt *et al.*, 1999; Vapnik, 1998). But all these algorithms are supplied with data in batch and thus involve a large amount of computation. Recently, various online SVM algorithms (Borders, 2005; Kivinen *et al.*, 2004; Lau and Wu, 2003) have been proposed to extend the SVM to the online setting.

An online SVM training algorithm (LASVM) that can be used on a large datasets. Borders (2005) presented a SVM algorithm called LASVM. It tolerates smaller main memory and has a faster training phase.

The standard online SVM algorithms are designed to handle binary problem. Due to the ever-changing experimental conditions, the model has to be updated continuously. The biologist hopes that after manually labeling some misclassified samples, the classifiers can be updated automatically and the resulting new model can be used to classify new examples. However, to apply online SVM to the task of cell phase identification, three factors have to be considered in advance. First, the datasets are critically imbalanced. The classification accuracy will undesirably bias toward the large classes. Second, some classes with fewer samples are more important than those with larger training samples. For example, prophase plays an important role in identifying the starting point of the mitosis process, but there are only about 140 examples of prophase in 200 frames of microscope images. Last, this is a multiclass classification problem.

2.5.2 Online support vector classifier Suppose we have one previously trained model and a set of new examples, i.e. a series of input vectors $\mathbf{x}_i \in \mathbf{R}^d$ $(i = 1, \ldots, l)$. The online learning will be discussed under the framework proposed by Shalev-Shwartz and Singer (2006).

In the online setting, on trial t, where $t \in [1, l]$, the online learning task can be regarded as solving the following optimization problem:

$$\min_{\omega \in R^d} P_t(\omega) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{t-1}\xi_i \tag{6}$$

with the first $t - 1$ examples being $\{x_1, y_1, \ldots, (x_{i-1}, y_{i-1})\}$.

Similarly, the dual objective function of $\min_{\omega \in R^d} P_t(\omega)$ is given by:

$$D_t(\boldsymbol{\alpha}) = \sum_{i=1}^{t-1} a_i - \frac{1}{2}\sum_{i=1}^{t-1}\sum_{j=1}^{t-1} a_i a_j y_i y_j \cdot K(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

From the definitions of $D_t(\boldsymbol{\alpha})$ and $D(\boldsymbol{\alpha})$, we can deduce that $D_t(\alpha_1, \ldots, \alpha_{t-1}) = D(\alpha_1, \ldots, \alpha_{t-1}, 0, \ldots 0)$. Thus, the online SVM can be treated as an incremental solver of the problem $\max_{\alpha \in [0,C]^m} D(\boldsymbol{\alpha})$ s.t.$\forall i > t, \alpha_i = 0$ (Shalev-Shwartz and Singer, 2006). It is worth noting that this problem can be solved in the sequential manner. For example, on trial t, $D(\boldsymbol{\alpha})$ only depends only on the first t observed variables. Intuitively, the larger the increase in the dual objective on each trail, the better the online SVC adjusts itself (Shalev-Shwartz and Singer, 2006).

Inspired by LASVM and the aforementioned framework, an online support vector is proposed for the problem of the cell phase identification. Three modifications are made. First, only the misclassified examples are used to update the model. Since new examples are critically imbalanced, the dominant class, i.e. interphase, will bring overwhelming information compared to other classes. Second, when trying to add a new example into the current support vector set, the coefficient α_i is initialized with different weights according to the importance of each class (Shalev-Shwartz and Singer, 2006; Wang *et al.*, 2005). Finally, the support vectors of the old model will be discarded once they become blatant non-support vectors during the optimization process. 'One-Verse-One' strategy is adopted to convert binary SVM to handle multiple classes (Vapnik, 1998).

We need to maintain four pieces of information: the set S of potential support vector indices, the set S_{old} of support vector indices that belong to the old model, the coefficients α_i of the old kernel expansion and the partial derivative g_i. The procedures INSERTION and UPDATING are two basic blocks of OSVC(Online Support Vector Classifier).

The basic idea of OSVC can be formulated as below. (1) Load the previously trained model, (2) the procedure INSERTION attempts to insert the misclassified example into the current kernel expansion and (3) the procedure UPDATING updates the model.

Algorithm (*Online Support Vector Classifier*)

```
(1) Initialization:
    Load the old support vector model.
        $S \leftarrow S_{old}; b \leftarrow b_{old}$
(2) Online Iteration:
    Set $\{\mathbf{x}_k, y_k\}$, for $k = 1, 2 \ldots l$
    for $k = 1, \ldots l$ do
    Obtain a new example $S_k = \{\mathbf{x}_k, y_k\}$
    Compute $\hat{y} = \sum_{i \in S} \alpha_i y_i K(\mathbf{x}_t, \mathbf{x}_i) + b$
    if example $\mathbf{x}_k$ is misclassified then
    INSERTION ($k$)
    UPDATING ($k$)
    end if
    end for
```

In INSERTION, the misclassified example $\mathbf{x}_k$, $k \in S$ is inserted into current support vector set. The coefficient α_k is assigned with the preset values C_P for positive example and C_n for negative example (Shalev-Shwartz and Singer, 2006; Wang *et al.*, 2005). The aims of assigning different weights are two fold. (1) The algorithm's performance is also impaired when the training sets with uneven class sizes are used. When the class of interest only has limited training samples, its prediction accuracy will undesirably decrease. To solve this problem, we can assign greater weight to the class of interest and thus improve its accuracy. (2) When the experimental condition changes greatly, the new examples can be assigned with larger weights to reflect the changes in the experiment. The online adaptivity of our algorithm can thus be improved. Then, a direction search will be performed to update the coefficients α_i and g_i.

```
INSERTION ($k$):
Obtain a new example $S_k = \{\mathbf{x}_k, y_k\}$
Compute $\hat{y} = \sum_{i \in S} \alpha_i K(\mathbf{x}_t, \mathbf{x}_i) + b$
if example $\mathbf{x}_k$ is misclassified, then
if $y_k = +1$ then
        $\alpha_k \leftarrow C_p, g_k \leftarrow y_k - \sum_{s \in S} \alpha_s y_s K_{ks}, S \leftarrow S \cup \{k\}$
        $i \leftarrow k, j \leftarrow \arg\min_{s \in S} g_s$   with $\alpha_s < C_n$
else
        $\alpha_k \leftarrow C_n, g_k \leftarrow y_k - \sum_{s \in S} \alpha_s y_s K_{ks}, S \leftarrow S \cup \{k\}$
        $i \leftarrow k, j \leftarrow \arg\max_{s \in S} g_s$ with $\alpha_s < C_p$
end if
if ($i$, $j$) is a $\tau$-violating pair
        $\lambda \leftarrow \min\{\frac{g_i - g_j}{K_{ii} + K_{jj} - 2K}, C_p - \alpha_i, C_n - \alpha_j\}$
        $\alpha_i \leftarrow \alpha_i + \lambda, \alpha_j \leftarrow \alpha_j - \lambda$
        $g_s \leftarrow g_s - \lambda(K_{is} - K_{js}) \forall s \in S$
end if
end if
```

The procedure UPDATING keeps on searching the τ-*violating* pair from the current support vector set and updating coefficients $\boldsymbol{\alpha}$ to increase the dual objective until there are no more such pairs. Analogous to REPROCESS in LASVM, the blatant non support vectors belonging to the old model will be removed (Borders, 2005).

```
UPDATING ($k$):
While there exists the $\tau$-violating pair ($i$, $j$) do
        $\lambda \leftarrow \min\{\frac{g_i - g_j}{K_{ii} + K_{jj} - 2K}, C_p - \alpha_i, C_n - \alpha_j\}$
        $\alpha_i \leftarrow \alpha_i + \lambda, \alpha_j \leftarrow \alpha_j - \lambda$
        $g_s \leftarrow g_s - \lambda(K_{is} - K_{js}) \forall s \in S$
        $i \leftarrow \arg\max_{s \in S} g_s$with $\alpha_s < C_p$
        $j \leftarrow \arg\min_{s \in S} g_s$with $\alpha_s < C_n$
for each $\alpha_S = 0$ and $s \in S_{old}$
if $y_s = -1$ and $g_s \geq g_i$, then $S = S - \{S\}$; end if
if, $y_s = +1$ and $g_s \leq g_i$ then $S = S - \{S\}$; end if
                $b \leftarrow (g_i + g_j)/2$
end for each
end while
```

3 RESULTS

3.1 Segmentation

To test the segmentation algorithm, 240 images are used. This generates a test set consisting of 18 683 nuclei. Three approaches are tested. (1) Simple watershed algorithm without fragments merging (Vincent and Soille, 1991). (2) Watershed and hybrid merging algorithm (Yan *et al.*, 2006). (3) The method proposed in this article. The hybrid merging algorithm can correctly identify 86% of the nuclei objects. Our proposed method, however, can correctly recognize 99% of the nuclei. Among the 18 683 nuclei, our method resulted in 35 nuclei that were over segmented and 154 nuclei that were under segmented. Most of the over segmentation is due to abnormal cell masses. The comparisons between these methods are given in Table 1. The details of segmentation validation are provided in the Supplementary Material.

3.2 Cell phase identification

We use sensitivity and specificity as the measurements of our experimental results. Suppose TP, TN, FP and FN stand for the number of true positive, true negative, false positive and false negative samples, respectively, after the completion of cell phase identification. Sensitivity is defined as sensitivity = TP/(TP + FN), and specificity is defined as specificity = TN/(TN + FP). In other words, sensitivity is a statistical measure of how well classified the positive cells are while specificity reflects the ability to identify negative cells correctly.

Table 1. Accuracy of nuclei recognition

Method	Under	Over	Correct
Watershed	187 (1%)	2242 (12%)	16 254 (87%)
Watershed + Hybrid	246 (1.32%)	747 (4%)	17 690 (94.69%)
Detection and watershed	154 (0.82%)	35 (0.19%)	18 494 (99%)

We can calculate the sensitivity and specificity for each class if we treat one class as positive and the others as negative.

Three ‘movies’ are manually labeled for testing the OSVC algorithm, which correspond to three successive experiments with different experimental conditions. The first two movies are untreated and the third is treated with taxol. Each movie contains 240 images. SVM (Vapnik, 1998) and LASVM (Borders, 2005) are used as baseline algorithms. The second and third movies are each separated into two halves. The first halves are used as training sets and the second as testing sets. In the testing set of second movie, there are 9249 interphase cells, 71 prophase cells, 183 metaphase cells and 180 anaphase cells. In the second half of third movie, there are 13 930 interphase, 48 prophase cells, 539 metaphase cells and 225 anaphase cells. For the untreated case, SVM (Vapnik, 1998) is trained with the cells in the first movie and the first half of the second movie. For the LASVM and online SVC, the ‘old’ model is generated with the first movie while the first half of second movie is used to update the ‘old’ model with the proposed online learning algorithm. For the treated case, the combination of the first movie and third movie is used. This is in contrast to the untreated case, which needs the first and the second movie. It is worth nothing that the OSVC algorithm can be updated continuously and do not need to wait until hand-labeling half of the movie and then updating the classifier. In other words, if the user meets one misclassified sample, he can update the classifier immediately by recording this new sample and discarding the outdated support vectors. It is for the convenience of comparison that use the first half as the training set, since standard SVM (Vapnik, 1998) is updated in a ‘‘batch’’ mode. It is the parameters obtained with cross-validation test are used for all algorithms: C is 0.707 and the gamma value for RBF kernel is 0.25. In OSVC, both weighted and non-weighted cases are tested.

Table 2. Sensitivity of cell phase identification for the untreated case

Phase	Inter (%)	Pro (%)	Meta (%)	Ana (%)
SVM	99.6	56.3	67.2	66.6
LASVM	99.8	45.1	77.2	82.0
OSVC (weighted)	94.9	87.3	66.1	84.0
OSVC	99.2	77.1	87.9	89.4

Table 3. Sensitivity of cell phase identification for the treated case

Phase	Inter (%)	Pro (%)	Meta (%)	Ana (%)
SVM	99.4	18.8	56.4	77.8
LASVM	98.5	22.9	75.1	86.7
OSVC (weighted)	90.9	77.0	79.2	77.3
OSVC	98.5	25.0	70.9	83.6

Table 4. Specificity of cell phase identification for the untreated case

Phase	Inter (%)	Pro	Meta (%)	Ana (%)
SVM	67.7	99.7	99.8	99.9
LASVM	80.7	99.9	99.8	99.9
OSVC (weighted)	95.9	99.6	99.8	99.8
OSVC	80.1	99.6	99.0	98.6

Table 5. Specificity of cell phase identification for the treated case

Phase	Inter (%)	Pro (%)	Meta (%)	Ana (%)
SVM	67.0	99.6	99.8	99.6
LASVM	80.6	99.9	99.8	99.8
OSVC (weighted)	95.9	99.6	99.8	99.8
OSVC	80.9	99.6	99.0	98.6

In weighted OSVC, the weights for interphase, prophase, metaphase and anaphase are 1, 30, 10 and 10, respectively. In Tables 2–5, the third rows of ‘OSVC weighted’ correspond to the weighted cases. In the non-weighted case, the C_p and C_n in UPDATING (k) are set to zero. Tables 2–5 show the sensitivity and specificity of both treated and untreated cases. Each line is an approach and each column is a phase. From Tables 2 and 3, the sensitivity of prophase with LASVM decreases while OSVC can increase the sensitivity greatly for both the untreated and treated case. In Table 5, the specificity of interphase can be improved significantly with weighted OSVC. The weights of OSVC can be tuned adaptively, which is useful when our focus is aimed at the classes having limited training samples. Here we give two general guidelines. (1) The more important cell types should be assigned greater weights. (2) At the initial stage, the weight for each class and the number of samples are in inverse proportion. For example, if the weight for prophase is 1 and the ratio between the numbers of prophase and metaphase is 10, then the weight for metaphase can be assigned 10.

4 DISCUSSION

Cells that undergo mitotic catastrophe will arrest, and eventually undergo apoptosis, before, during, or just after the metaphase-anaphase transition. Thus, we will detect these cells as frozen in metaphase or anaphase from image to image, allowing us to classify them as arrested metaphase.

Nuclei segmentation is an essential part of the whole system. To separate the dark nuclei from the background, a data-driven background correction algorithm is used as an adaptive thresholding method. In order to reduce the over-segmentation and under-segmentation problems, we first combine the shape information and intensity information of the nuclei together, and then the linear scale space theory is used to generate a local maximum that represents the position of one nucleus. Finally, the seeded watershed algorithm, in which the detected central points (positions of nuclei) are used as 'seeds', is used to segment the nuclei. The experimental results show that the proposed segmentation algorithm works well.

It is more challenging to deal with more irregular cell types, like U87-MG, which requires more robust segmentation method. But our system can work well even when such cells show up. The main aim of this research is to estimate the changes in the cell cycle before and after drug use. Prophase, interphase, metaphase and anaphase are the four cells types of our concern. The other cell types, like U87-MG, G0/G1, etc. seldom appear in our experiments and their lasting time are very short. So, we can use the context information to correct the errors caused by the incorrect segmentation. In addition, a more robust algorithm usually incurs more computational overhead, which is another important factor that must be considered in a real-time system. Therefore, we focus on improving the classification accuracy of these four cell types instead of employing a more complex segmentation method.

To solve the problem caused by imbalanced datasets, the weighted OSVC is proposed to improve the prediction accuracy of interest. For the untreated case, both the weighted and non-weighted OSVC perform better than SVM and LASVM. The weighted version can be applied when our focus is one particular class. In the treated case, only the weighted SVM can effectively improve the prediction accuracy of prophase due to the dramatic morphological changes of the treated cells. However, the weights for each class should be tuned to meet the real experimental requirements.

It is worth noting that the problems with unequal distribution of training examples can be solved by the 'weighted' SVM (Eitrich and Lang, 2006; Vapnik, 1998), which trains the SVMs by assigning the training samples with different cost weights according to class size. In the 'weighted' SVM, the prediction accuracy of prophase can be increased by 10–20% at the expense of slightly decrease of the classes with large samples using the reduced features. The weights for interphase, prophase, metaphase and anaphase are 1, 10, 10 and 10.

5 CONCLUSION AND FUTURE WORKS

In this article, a novel cell nuclei detection algorithm is proposed to work with watershed algorithms for cell nuclei segmentation of imaging for drug discovery and quantitative biology studies. Based on the segmentation results, an OSVC algorithm is used to classify the nuclei into different phases and the performance has been validated for the online cell phase identification. Future work is to construct a database from vast amounts of images of cancer cell lines under different drug perturbation conditions. Following this, the influence of various drug components in mitotic process will be analyzed to find the key anti-mitotic cancer drug components.

ACKNOWLEDGEMENT

This research is funded by an NIH R01 LM008696 Grant.

Conflict of Interest: none declared.

REFERENCES

Anderson,H.J. *et al.* (2003) Inhibitors of the G2 DNA damage checkpoint and their potential for cancer therapy. *Prog. Cell Cycle Res.*, **5**, 423–430.

Baguley,B.C. and Marshall,E.S. (2004) In vitro modeling of human tumour behaviour in drug discovery programmes. *Eur. J. Cancer*, **40**, 794–801.

Boland,M.V. and Murphy,R.F. (2001) A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope image of HeLa cells. *Bioinformatics*, **17**, 1213–1223.

Boland,M.V. *et al.* (1998) Automated recognition of pattern characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, **33**, 366–375.

Borders,A. (2005) Fast kernel classifiers with online and active learning. *J. Mach. Lean. Res.*, **6**, 1579–1619.

Breu,H. *et al.* (1995) Linear time Euclidean distance transform algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**, 529–533.

Charles,Y.T. *et al.* (2007) A Support Vector Machine Classifier for recognizing mitotic subphases using high-content screening data. *J. Biomol. Screen.*, **12**, 490–496.

Chen,S.C. and Murphy,R.F. (2006) A graphical model approach to automated classification of protein subcellular location patterns in multi-cell images. *BMC Bioinformatics*, **7**, 90.

Chen,X. *et al.* (2006) Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Trans. Biomed. Eng.*, **53**, 762–766.

Cortes,C. and Vapnik,V. (1995) Support vector networks. *Mach. Learn.*, **20**, 273–297.

Debes,J.D. *et al.* (2005) P300 Modulates Nuclear Morphology in Prostate Cancer. *Cancer Res.*, **65**, 708–712.

Dixon,H. *et al.* (2002) Therapeutic exploitation of checkpoint defects in cancer cells lacking p53 function. *Cell Cycle*, **1**, 362–368.

Eitrich,T. and Lang,B. (2006) Efficient optimization of support vector machine learning parameters for unbalanced datasets. *J. Comput. Appl. Math.*, **196**, 377–427.

Freund,Y. and Schapire,R.E. (1999) Large margin classification using the perceptron algorithm. *Mach. Learn.*, **37**, 277–296.

Gallardo,G. *et al.* (2004) Mitotic cell recognition with hidden markov models. *Med. Imaging SPIE*, **5367**, 661–668.

Guyon,I. *et al.*utomatic capacity tuning of very large VC-dimension classifiers. In *Advances in Neural Information Processing Systems*. Morgan Kaufmann, CA, pp. 5–155.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Haralick,R.M. *et al.* (1973) Textural features for image classification. *IEEE Trans. SMC.*, **3**, 610–621.

Harder,N. *et al.* (2006) Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. In Larsen,R. *et al.* (eds.) *Proceedings of the MICCAI'06*. Springer Berlin/Heidelberg, Copenhagen, DK, pp. 840–848.

Huang,K. *et al.* (2003) Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope image. *Proc. SPIE*, **4962**, 307–318.

Kivinen,J. *et al.* (2004) Online learning with Kernels. *IEEE Trans. Signal Process.*, **52**, 2165–2176.

Lau,K.W. and Wu,Q.H. (2003) Online training of support vector classifier. *Pattern Recognit.*, **36**, 1913–1920.

Li,G.Z. *et al.* (2004) Feature selection for multi-class problems using support vector machines. *Lect. Notes Comput. Sci.*, **3157**, 292–300.

Lin,G. *et al.* (2003) A hybrid 3-D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. *Cytometry A*, **56**, 23–36.

Lindblad,J. *et al.* (2004) Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation. *Cytometry A*, **57**, 22–33.

Manjunath,B.S. and MA,W.Y. (1996) Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI (PAMI-Special issuer on Digital Libraries)*, **18**, 837–842.

Murphy,R.F. *et al.* (2003) Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *J. VLSI Signal Process.*, **35**, 311–321.

Platt,J. *et al*. (1999) Fast training of support vector machines using sequential minimal optimization. In Schölkopf,B. *et al*. (eds.) *Advances in Kernel Methods-Support Vector Learning*. MIT Press, MA, pp. 185–208.

Rosenblatt,F. (1958) The perceptron: a probability model for information storage and organization in the brain. *Psychol. Rev*., **6**, 386–408.

Shalev-Shwartz,S. and Singer,Y. (2006) Online learning meets optimization in the dual. In Carbonell,J.G. and Siekmann,J. (eds.) *Proceedings of the 19th Annual Conference on Learning Theory*. Springer Berlin/Heidelberg, Pittsburgh, PA, pp. 423–437.

Steger,C. (1998) An unbiased detector of curvilinear structures. *IEEE Trans. Pattern Anal*., **20**, 113–125.

Stern,H.M. *et al*. (2005) Small molecules that delay S phase suppress a zebrafish bmyb mutant. *Nat. Chem. Biol*., **1**, 366–370.

Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley & Sons, NY.

Vapnik,V. and Lerner,A. (1963) Pattern recognition using generalized portrait method. *Automat. Rem. Contr*., **24**, 774–780.

Vincent,L. and Soille,P. (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. PAMI*, **13**, 583–598.

Wahlby,C. *et al*. (2002) Algorithms for cytoplasm segmentation of fluorescence labelled cells. *Anal. Cell Pathol*., **24**, 101–111.

Wang,M. *et al*. (2005) Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. *Protein Eng. Des. Sel*., **17**, 509–516.

Wang,M. *et al*. (2007) Context based mixture model for cell phase identification in automated fluorescence microscopy. *BMC Bioinformatics*, **8**, 32.

Xu,C. and Prince,J. (1998) Snakes, shapes, and gradient vector flow. *IEEE Trans. Image Process*., **7**, 359–369.

Yan,J. *et al*. (2006) An efficient system for optical microscopy cell image segmentation, tracking and cell phase identification. In *Proceedings of IEEE ICIP*. Atlanta, GA, pp. 1536–1537.

Zhou,X. and Wong,S.T.C. (2006) Informatics challenges of high-throughput microscopy. *IEEE Signal Process. Mag*., **23**, 63–72.