Journal of Microscopy, Vol. 226, Pt 2 May 2007, pp. 121–132
Received 21 August 2006; accepted 3 January 2007

An automated feedback system with the hybrid model of scoring and classification for solving over-segmentation problems in RNAi high content screening

F. LI*†, X. ZHOU*‡, J. MA† & STEPHEN T. C. WONG*‡
** Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair, Harvard Medical School, 3rd floor, 1249 Boylston, Boston, MA 02115, U.S.A.*
†Department of Information Science, School of Mathematical Sciences, and LMAM, Peking University, Beijing 100871, China
‡Department of Radiology, Brigham and Women's Hospital, Boston, MA 02115, U.S.A.

Key words. Automated cell segmentation, feature selection, high content screening, model-based merging, RNAi.

Summary

Background: High content screening (HCS) via automated fluorescence microscopy is a powerful technology for generating cellular images that are rich in phenotypic information. RNA interference is a revolutionary approach for silencing gene expression and has become an important method for studying genes through RNA interference-induced cellular phenotype analysis. The convergence of the two technologies has led to large-scale, image-based studies of cellular phenotypes under systematic perturbations of RNA interference. However, existing high content screening image analysis tools are inadequate to extract content regarding cell morphology from the complex images, thus they limit the potential of genome-wide RNA interference high content screening screening for simple marker readouts. In particular, over-segmentation is one of the persistent problems of cell segmentation; this paper describes a new method to alleviate this problem.

Methods: To solve the issue of over-segmentation, we propose a novel feedback system with a hybrid model for automated cell segmentation of images from high content screening. A Hybrid learning model is developed based on three scoring models to capture specific characteristics of over-segmented cells. Dead nuclei are also removed through a statistical model.

Results: Experimental validation showed that the proposed method had 93.7% sensitivity and 94.23% specificity. When applied to a set of images of F-actin-stained Drosophila cells, 91.3% of over-segmented cells were detected and only 2.8% were under-segmented.

Conclusions: The proposed feedback system significantly reduces over-segmentation of cell bodies caused by over-segmented nuclei, dead nuclei, and dividing cells. This system can be used in the automated analysis system of high content screening images.

Correspondence to: Stephen T.C. Wong. Tel: 617-525-7595; fax: 617-525-6220; e-mail: stephen_wong@hms.harvard.edu

1. Introduction

High content screening (HCS) by automated fluorescence microscopy is becoming an important and widely used research tool to assist scientists in understanding complex cellular processes, such as mitosis and apoptosis, as well as in disease diagnosis and prognosis, drug target validation and compound-led selection (Perlman *et al.*, 2004; Zhou & Wong, 2006a,b). Meanwhile, RNA interference (RNAi) is a revolutionary approach for silencing gene expression and has become an important method for analyzing gene function. The convergence of the two technologies has led to large-scale, image-based studies of cellular phenotypes by systematic perturbation using RNAi. Indeed, cellular images generated by the RNAi-HCS technology are relatively new to image analysis and pattern recognition communities such that existing HCS image analysis tools are inadequate to delineate and extract the rich morphologic content of cellular phenotypes from the complex images, a problem that significantly restricts the potential of HCS in systems biology and drug discovery. However, since it is time-consuming and impractical to segment cells manually for vast amounts of image data sets generated in HCS studies, the availability of fully automated cell image segmentation and quantification system is critical to the success of HCS.

Using images acquired by automated microscopy, biologists visualize phenotypic changes resulting from reverse-functional analysis by the treatment of Drosophila cells in culture with gene specific double-stranded RNAs (dsRNAs), which 'knocks-out' target gene function by RNAi technology (Boutros *et al.*, 2004). Even for a small scale RNAi study by manual analysis, a wide range of phenotypes with affected cytoskeletal organization and cell shape was observed (Kiger *et al.*, 2003). However, in genome-wide RNAi-HCS studies, there are more than 21 000 gene-specific dsRNAs, resulting in hundreds of thousands of images in a single experiment (see Section 2.1). Consequently, it is impossible to characterize and quantitate the morphological phenotypes manually. A fully automated, robust cell image analysis system is needed.

A number of automated methods for segmentation of nuclei and cell bodies are available. These methods can be generally classified into three categories: deformable model, Voronoi diagram, and watershed. Segmentation algorithms using deformable models are popular in which cell contours evolve under the direction of internal and external forces from initial contours until they reach cell boundaries. Two such deformable model methods, snake and level sets, are widely used to segment three-channel images (Kass *et al.*, 1987; Malladi *et al.*, 1995; Klemencic *et al.*, 1998; Chan & Vese, 2001; Xiong *et al.*, 2005). In these methods, however, edge detection is a function of the image gradient, which usually results in edge leaking, and the correct segmentation result closely depends on the initialization of the contours, which is also difficult. Moreover, these methods are known to be computationally expensive (see Xiong *et al.*, 2005). The general Voronoi diagram method only can detect the approximate position and region of cells (Morelock *et al.*, 2005). A novel variation of the Voronoi diagram method defines more accurate cell boundaries (Jones *et al.*, 2005); however, this method degenerates into the general Voronoi diagram method when image noise increases. The Voronoi diagram methods also require the initial positions of the cells. Finally, the watershed method and its variations (Beucher, 1992; Lin *et al.*, 2003; Vincent & Soille, 1991) suffer a drawback of over-segmentation. Although rule-based merging methods (Adiga & Chaudhuri, 2001; Wahlby *et al.*, 2002), e.g. size or integrated pixel intensity, can be used to reduce the over-segmentation, it is difficult to define reliable rules, and these heuristic rules are often prone to error in processing complicated cell images. Consequently, seeded watershed is commonly used to reduce the over-segmentation (Vincent & Soille, 1991; Beucher, 1992; Lin *et al.*, 2003, 2005; Lindblad *et al.*, 2004). Segmenting nuclei is relatively easy because of their regular shapes and high intensity relative to background signals. So the position and contour information of the nuclei are widely used in aforementioned methods: the initial contouring in the deformable based methods, the cell positions in the Voronoi diagram based methods, and the 'seeds' in the watershed methods. Cellomics (Kapur, 2001), Q3DM (Morelock *et al.*, 2005), GE-InCell Analyzer (Lindblad *et al.*, 2004) and CellProfiler (Jones *et al.*, 2005) are commercially or publicly available software for cellular image analysis, and they employ either Voronoi diagram (propagation) or watershed methods for cell segmentation.

All the segmentation methods discussed above are dependent on the segmented nuclei for three-channel based analysis, a strategy with serious drawbacks. The drawbacks lay on the following facts. To visualize nuclei, cells are stained with the ultraviolet-fluorescing, DNA-binding molecule 4′,6-diamidino-2-phenylindole (DAPI). The inherent properties of cells and the DAPI stain lead to over-segmentation. Especially, when cells are actively dividing, DNA condenses and separates into two regions, two nuclei visually, within a single cell. Consequently, the two nuclei within one cell would result in over-segmentation (Fig. 1A). Additionally, erroneous over-segmentation of nuclei also causes over-segmentation of the cell directly (Fig. 1B). Finally, there are occasionally some nuclei that have little or no cytoplasm (based on staining captured in the F-actin channel). These nuclei, which we refer to as dead nuclei, also cause over-segmentation of the cell cytoplasm (Fig. 1C). Lin *et al.* (2003, 2005) proposed a machine learning method to reduce over-segmentation in *nuclear* segmentation, however, it cannot efficiently solve the problem of over-segmentation of *cell cytoplasm* as the large number of irregular shapes and topological structures found in cell cytoplasm images generated by RNAi screening.

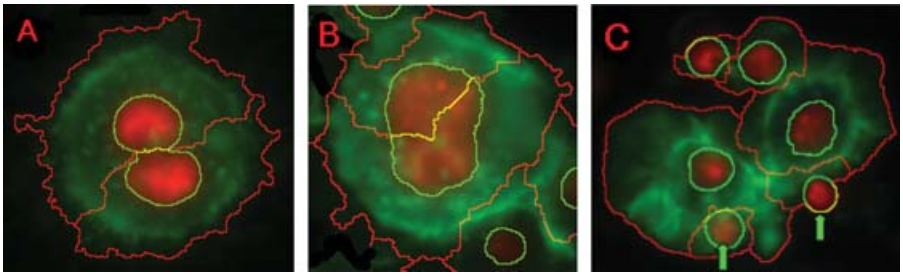

Fig. 1. Multiple factors contribute to over-segmentation of cells. Green is F-actin staining, which visualizes the cytoskeleton; red is DAPI staining, which visualizes DNA within the nucleus. A: An actively dividing cell, which has two DAPI-stained regions, is over-segmented. B: Over-segmented nucleus results in the over-segmentation of cell directly. C: Dead nuclei (arrows) result in over-segmentation of cells.

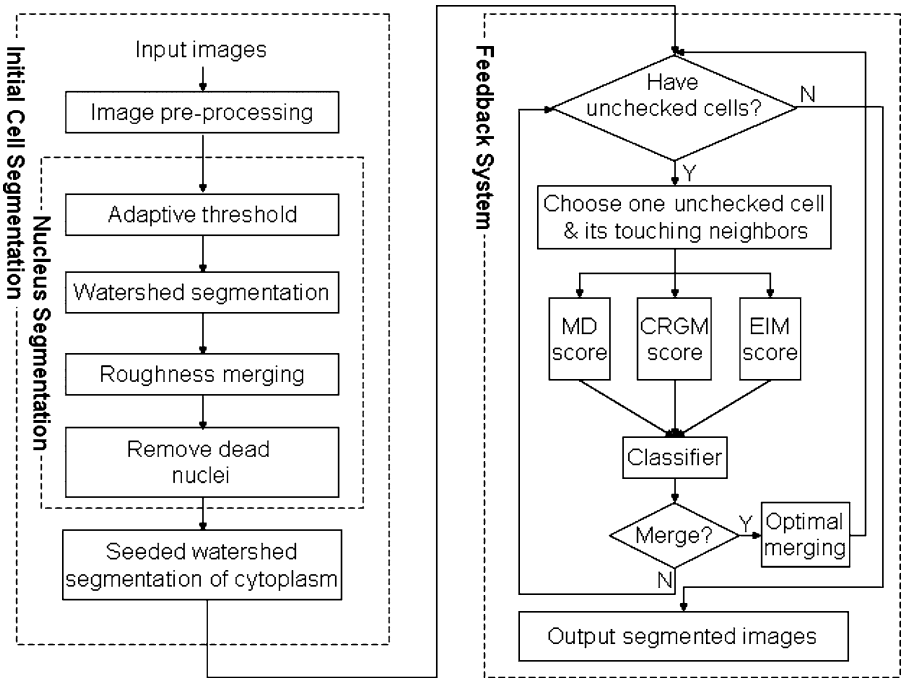

Fig. 2. An overview of the flowchart of the entire segmentation system for RNAi high content screening image analysis.

The goal of the present work is to build a fully automated RNAi cell image segmentation system that can be used to quantify the behaviours of the cells for the gene functions' research. The specific focus of the present work is to reduce the over-segmentation of cell cytoplasm in HCS images, which cannot be solved effectively by the existing methods. To achieve this task, we first filter out the dead nuclei via a statistical approach and then reduce the over-segmentation of the nuclei. In particular, we build a novel feedback system in which three scoring models are defined carefully to capture the specific differences between the over- and well-segmented cells. Using a well known statistical classifier, quadratic discriminant analysis (QDA), the over-segmented cells are then identified and merged. Figure 2 illustrates the flowchart of the proposed system. The system is composed of two major modules: initial cell segmentation and feedback merging system. The initial cell segmentation consists of six sub-modules: (1) image noise is suppressed in the image pre-processing modules; (2) the DAPI signal (nucleus) is separated from the background of the DNA image using adaptive thresholding; (3) nuclear regions are segmented by watershed segmentation; (4) over-segmentation of nuclei is reduced via roughness merging method; (5) dead nuclei are filtered out and (6) cell bodies are segmented via a seeded watershed model using the separated nuclei as seeds. The feedback system entails three steps. First, three scoring models are defined to capture specific characteristics of the over-segmented cells. Secondly, a classifier maps the three scores into a merging decision. Finally, over-segmented cells are detected and merged according to merging decisions.

2. Materials and methods

2.1. *RNAi-HCS images*

A genome-wide screen examines more than 21,000 dsRNAs, specific to predicted Drosophila genes. The dsRNAs are robotically arrayed in 384-well plates. Drosophila cells are plated and taken up the dsRNA from culture media so that the desired assay is performed. After the desired incubation time with the dsRNA, cells are fixed, stained and imaged by automated microscopy. For each dsRNA treatment, three sites are imaged, and for each site as many as three channels of different cellular markers are acquired. Thus, a single replicate will generate ~200 000 images with the size of 1280 × 1024 pixels. In the present study, a cell-based assay for Rho GTPase activity was developed using the Drosophila Kc167 embryonic cell line, which is described elsewhere (Kiger *et al.*, 2003). Three distinct cellular phenotypes are observed: normal, spiky and ruffling (Fig. 3).

2.2. *Image pre-processing*

Before segmenting an image, it is necessary to perform pre-processing because the noises, artefacts, uneven illumination, and striped patterns would degrade image quality (Wahlby *et al.*, 2002; Lin *et al.*, 2003; Lindblad *et al.*, 2004). To remove the noises and other artefacts without blurring the edges, median filtering is commonly used (Wahlby *et al.*, 2002; Lin *et al.*, 2003). For uneven illumination and striped patterns, a data-driven background algorithm (Wahlby *et al.*, 2002;

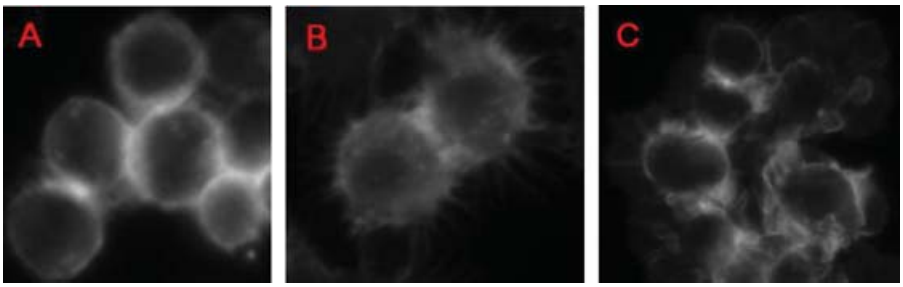

Fig. 3. Three distinct cellular phenotypes are observed. A: Normal. B: Spiky. C: Ruffling. All panels are stained with TRITC-phalloidin to visualize F-actin.

Lindblad *et al.*, 2004) is employed to correct the degradation of the images. The algorithm makes use of the cubic B-splines, which have good features of continuousness and smoothness to estimate the background iteratively, and the foreground pixels are detected gradually each time by subtracting the estimated background from the original image. The convergence of this algorithm is fast, and, after reducing the influences of uneven illumination and striped patterns, the resulting image has better quality also.

2.3. Initial cell image segmentation

Nuclear segmentation. First, the nuclei are separated from the background by using the adaptive threshold method (Otsu, 1978; Sahoo *et al.*, 1988). Next, watershed algorithm is applied to the distance-transformed image to separate the nuclear clusters. Instead of using the model-based merging algorithm described in (Lin *et al.*, 2003, 2005), a simple and efficient roughness-merging algorithm is adopted to reduce the over-segmentation of nucleus (Chen *et al.*, 2006). As can be seen in Figs 4A and B, the number of over-segmented nuclei is reduced significantly after roughness merging. Figure 4C is the manual segmentation result. Morphological open operation (Lin *et al.*, 2003) is then applied to smooth the nuclear boundaries.

Removal of dead nuclei. Certain drosophila cells died during screening. When cells are dying, their cytoplasm gradually disappears, and at the same time, the volume of the nuclei also shrinks. Thus, it causes the stain dye to sustain a high concentration. As a result, dead nuclei are clearly distinguished by their brighter intensity and smaller size than the living nuclei (Fig. 5A). The cell cytoplasm corresponding to these nuclei, as observed in the F-actin channel, has almost disappeared (Fig. 5B). The dead cells are not considered in the RNAi analysis; however, when dead nuclei are covered by the cytoplasm of living cells, it will cause over-segmentation of cell cytoplasm (see Fig. 1B). So, all the dead nuclei should be removed (Fig. 5C).

Detection of dead nuclei is a classification problem. First, specific features that can be used to describe the differences between the dead nuclei and the living nuclei are needed. For complex classification problems, (Zernike, 1934; Haralick, 1979; Cohen *et al.*, 1992; Manjunath and Ma, 1996; Wang *et al.*, 2006) extracted many features for every nuclei, out of which an approximately optimal subset of features was found using automated feature selection algorithms (Siedlecki and Sklansky, 1989; Pudil *et al.*, 1994; Jain & Zongker, 1997). We noticed that the differences between dead nuclei and living nuclei are distinct, for example, the dead nuclei have smaller size and higher intensity than the living cells. Thus, we directly extract the size, intensity and the standard intensity deviation of the nuclei as features to distinguish the dead nuclei from the living nuclei.

A well tested and widely used statistical QDA classifier (Duda and Hart, 1973; Lindblad *et al.*, 2004) is employed for classification. The QDA classifier makes use of two Gaussian distributions to fit the two clusters: dead and living nuclei. Then, Bayes' rule was used to calculate the posterior probabilities, and the maximum of the posterior probabilities was used to decide a nucleus' class. In our study, a training data set including 108 dead nuclei and 1, 243 living nuclei

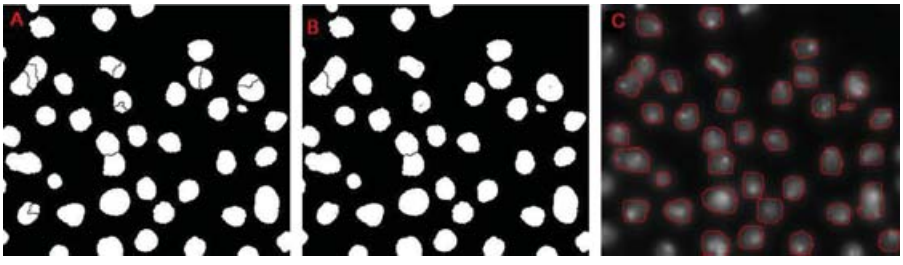

Fig. 4. Comparison of results before and after roughness merging. A: Result before the roughness merging step. B: Result after the roughness merging step. C: Manual segmentation result.

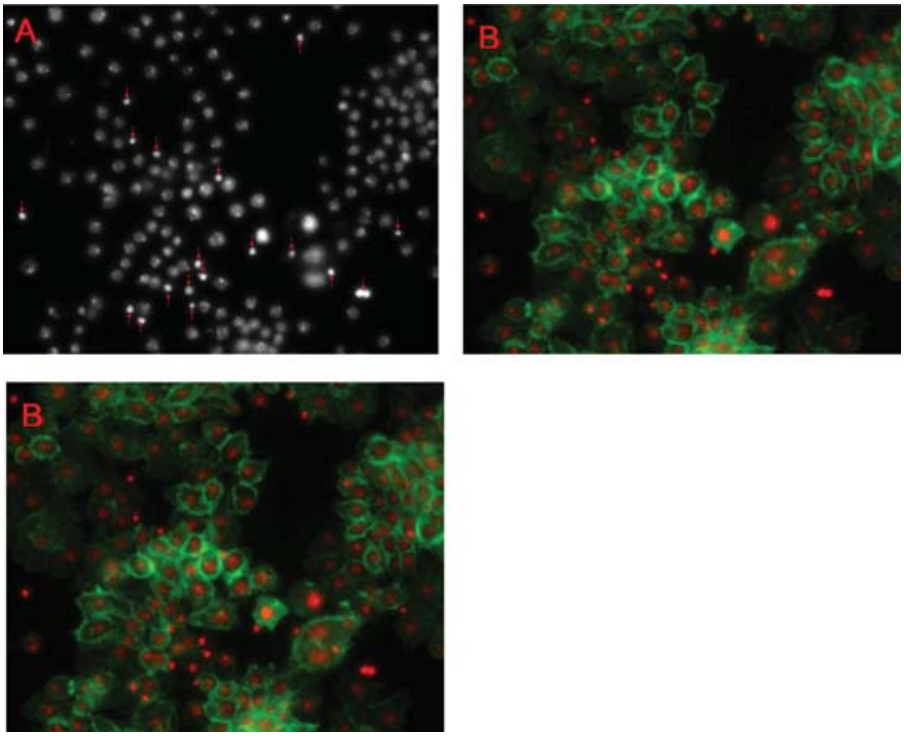

Fig. 5. Illustration of the procedure of removing the dead nuclei. A: The original nuclei image in the DNA channel with dead nuclei (red arrows). B: The colour image derived from the original DNA channel (red) and the F-actin channel (red). C: The colour image after removing the dead nuclei.

were collected to estimate the parameters in the QDA classifier. The QDA classifier is detailed in Section 2.4.4. After this step, most of dead nuclei were filtered out. Experimental result, see Section 3.1, shows that the sensitivity and specificity levels reach 95.83% and 95.47%, respectively.

Initial cell body segmentation. After segmenting nuclei, the cell cytoplasm can be separated via a seeded watershed method. First, fuzzy c-means threshold algorithm (Pham *et al.*, 2004; Zhou *et al.*, 2005) is applied to the F-actin channel to separate the cell cytoplasm from the background. Then, we apply the seeded watershed algorithm to the F-actin images directly to separate the touching cells since in the F-actin images, the skeletons (edges) of cells, always have higher intensity than the interior of the cells, so the watersheds will be built on the desired edges. The other reason is that the intensity of the cytoplasm often has a non-uniform variation inside the cells, so the gradient image cannot describe the boundary well.

2.4. *Feedback merging system*

To deal with the over-segmentation of cell cytoplasm, we developed a feedback system based on a hybrid model of scoring and classification. Three different scoring models are used to describe the characteristics of the over-segmented cells: Mahalanobis distance (MD) model, centre region gradient (CRG) model, and edge intensity (EI) model. We built a classifier to model the scores, $S = \{S_{MD}, S_{CRG}, S_{EI}\}$, derived from the three scoring models. The output y determined whether or not the two cells should be merged into a single cell with two nuclei. Here y takes 0 (not merging) or 1 (merging). Mathematically, the model can be given by:

$$y = f(S) = f(S_{MD}, S_{CRG}, S_{EI}) \quad (1)$$

In the following sections, we describe the three models used to generate scores and the classifying model. For notational convenience, let $c1$ and $c2$ denote two touching cells; e is their common edge. Cell c is the cell after merging cell $c1$ and $c2$ whereas $n1$ and $n2$ are nuclei corresponding to cells $c1$ and $c2$. On the other hand, $o1$, $o2$, and $r1$, $r2$ are the centroids and radii of nuclei of $n1$ and $n2$, respectively.

2.4.1. Mahalanobis distance model. We reason that there would be measurable differences between over-segmented cells and appropriately segmented cells. Two statistical models, the probability density function (PDF) model [14,16] and the MD model (Wahlby *et al.*, 2002), are widely used to measure the differences between two objects. Here, we use the MD model because the value range of the PDF model is too small for accurate analysis.

Feature extraction & automatic feature selection for MD model. A training data set, including 300 intact normal cells, 200 intact spiky cells, 200 intact ruffling cells, and 100 partial cells (cells that were identified manually as over-segmented), is collected from the initial segmentation result first. Next,

211 features (Wang *et al.*, 2006) are extracted to differentiate the geometric properties and appearances of three phenotypes and partial cells. Generally, the features are classified into five categories: three kinds of general features, including wavelet features (Cohen *et al.*, 1992; Manjunath & Ma, 1996), Zernike moments features (Zernike, 1934), and Haralick features (Haralick, 1979), and two kinds of specific shape descriptor features (Wang *et al.*, 2006). Then, the sequential floating forward selection algorithm (SFFS) (Pudil *et al.*, 1994) is adopted to select a subset of features to reduce the influence of over-training and to improve the ability of generalization of the classifier on new data sets. Linearly dependent or near-linearly dependent features were removed based on Pearson's linear correlation coefficients to avoid a singular matrix in the MD model. Finally, we select three subsets of features that distinguish partial cells from the intact normal, spiky and ruffling cells.

MD model and MD score. We derive a score, denoted as S^c_{MD}, from the MD models. For each cell, there were three MD models defined as:

$$d_N^2 = (\vec{x} - \vec{x}_N)' \Sigma_N^{-1} (\vec{x} - \vec{x}_N) \tag{2}$$

where $\vec{x}$ is the feature vector of input cell, $\vec{x}_N$ is the mean value of normal sample set and Σ_N^{-1} are normal sample covariance matrix in the given features space. Similarly, the MD models for spiky and ruffling phenotypes are constructed in the same way, by replacing the sample mean value vector $\vec{x}_N$ and sample covariance matrix Σ_N^{-1} with $\vec{x}_S$, $\vec{x}_R$ and Σ_S^{-1}, Σ_R^{-1}, respectively. Thus, for one cell c, three MD distances are calculated, $\{d_N^c, d_S^c, d_R^c\}$, describing the difference between the input cell c and the mean value of normal, spiky, and ruffling samples, $\vec{x}_N, \vec{x}_S, \vec{x}_R\}$.

Specifically, three MD distances are obtained for cells $c1$ and $c2$: $\{d_N^{ci}, d_S^{ci}, d_R^{ci}\}, i = 1, 2$ and cell c:$\{d_N^c, d_S^c, d_R^c\}$ (see Fig. 6). We define the MD scores of cells $c1$ and $c2$ as follows:

$$S^c_{\mathrm{MD}} = \frac{(D^{c1} + D^{c2})}{2 \times D^c}, \tag{3}$$

where

$$D^{ci} = \min\left(d_N^{ci}, d_S^{ci}, d_R^{ci}\right), i = 1, 2, \text{and} D^c = \min\left(d_N^c, d_S^c, d_R^c\right).$$

The minimum value of the three MDs reflects the difference between the input cell and its nearest phenotype (represented by the sample mean value vector) in the given features space. Thus, if the merging cell c is more similar to one phenotype than cells $c1$ and $c2$, D^cshould be less than D^{c1} and D^{c2}. It follows that a higher value of $S^c{}_{\mathrm{MD}}$ results in a greater probability of merging the two close-lying cells.

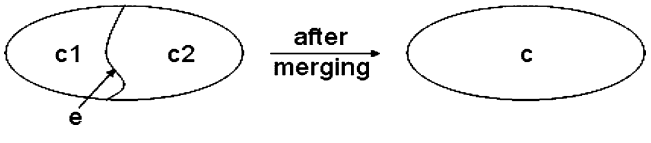

Fig. 6. Cells model in computing the MD score. $c1$ and $c2$ are two touching cells and $c$is the cell after merging cell $c1$ and $c2$.

2.4.2. Centre region gradient model We also take advantage of differences in the intensity patterns between cells to determine if a cell is over-segmented. We observed that the intensity of a region inside a cell is relatively flat (Fig. 7, left two columns), whereas an increase in the form of a ridge appears between two cells (Fig. 7, right two columns). These observations can be easily expressed by intensity gradients. We reason that if a cell is over-segmented, we will not find a significant intensity ridge between the two touching cells. We construct the CRG model using intensity gradient information between two neighbouring cells.

Cropping centre region. To capture the gradient information between two neighbouring cells accurately, a centre region R is cropped first (red region in Fig. 8). The long axis a is the distance between the two nuclear centroids, $a = d(o_1, o_2)$, and the short axis b is the average of the two nuclear diameters, $b = (r_1 + r_2)$. The centre region R has three important features. First, since nuclei always reside within cells, R is located by default within the two cells, thereby eliminating the noise of the true boundaries of the two cells. Second, since the common edge between the two cells must cross the region between two nuclei, edge information is captured within this rectangle region, R. Third, as the orientation of the ridge is roughly parallel to the short axis b, the length of the ridge, denoted as N_e, is proportional to the length of short axis b. This relationship offers a good criterion for the following definition of the score of the CRG.

CRG model and CRG score. After cropping the centre region R, the CRG model is defined as follows:

$$S^c_{\mathrm{CRG}} = \frac{C \times b}{\left(\sum_{(x,y)\in R} q(G(x, y) - T)\right) + 1}, \tag{4}$$

where S^c_{CRG} is the CRG score, b is the length of the short axis of R, and c is a constant whose value is used as a parameter to control the value range of S^c_{CRG}; this does not influence the final merging decision significantly. In our studies, we set $C = 15$ to make most of the partial cells' CRG score, $S^c{}_{\mathrm{CRG}}$, more than one, and that of intact touching cells less than one. $G(x, y)$ is the gradient value of pixel (x, y), and T is an edge threshold that determines whether pixel (x, y) is a ridge pixel or not, and T is set automatically by applying the Otsu' threshold method on the gradient image. $q(x)$ is an indictor function which takes a value of 1 when $x \geq 0$, and 0 otherwise. Therefore, the CRG score $S^c{}_{\mathrm{CRG}}$ indicates the nature of the edge between two touching cells: a higher value of $S^c{}_{\mathrm{CRG}}$ indicates that there is a weak edge and the two cells are likely to be part of an over-segmented cell, whereas a lower value of $S^c{}_{\mathrm{CRG}}$ means that there is a true,

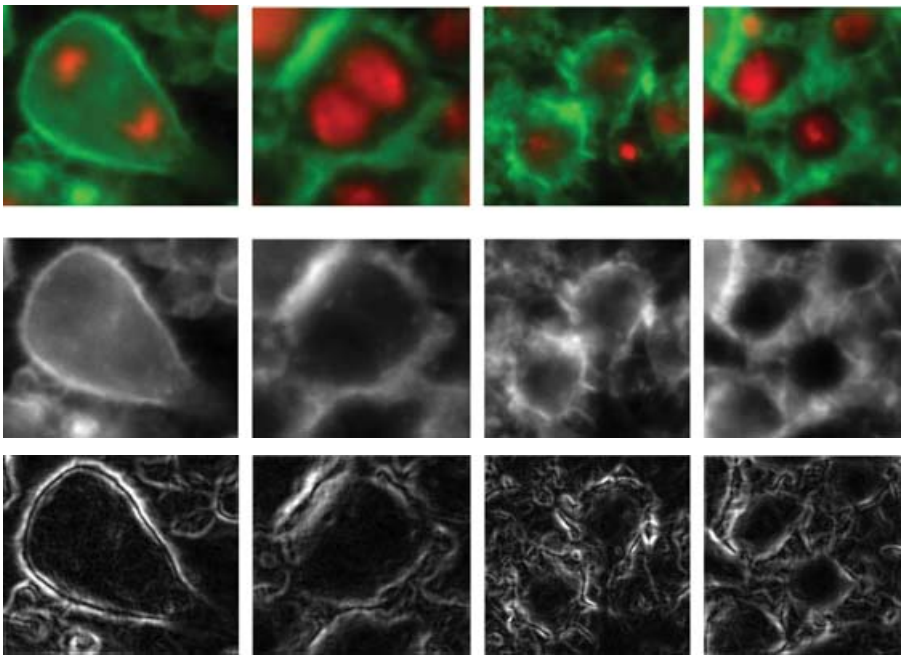

Fig. 7. Illustration of the ridge (gradient) information inside and between cells. The top row are the colour images derived from the original DNA channel and the actin channel, with the red objects are nuclei and the green material cytoplasm. The middle row shows the actin channel images. The bottom row displays the gradient images.

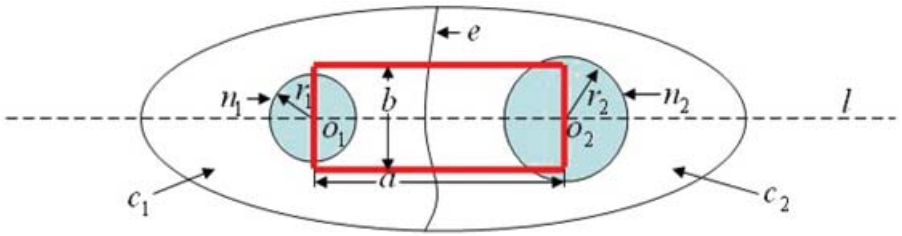

Fig. 8. Illustration of the cropping of centre region in the CRG model.

strong edge, indicating the two touching cells are segmented appropriately.

2.4.3. Edge intensity model. As mentioned above, the interior intensity of the cell is lower than its boundaries in the F-actin images, and the seeded watershed algorithm is applied to the intensity image directly. Therefore, if a cell is over-segmented, the common edges between the partial cells have lower intensity than the true edges, as can be seen in Fig. 9A, and the common edge between two intact cells shown in Fig. 9B.

EI model and EI score. Based on the above observation, the EI model is built as following:

$$S_{\mathrm{EI}}^{c} = \frac{\min\{\mathrm{ave}(I_{b1}), \mathrm{ave}(I_{b2})\}}{\mathrm{ave}(I_e)}, \qquad (5)$$

where S_{EI} is the EI score and $\mathrm{ave}(I_{b1})$, $\mathrm{ave}(I_{b2})$ and $\mathrm{ave}(I_e)$ are the average intensities of boundaries $b1$, $b2$ and e, respectively.

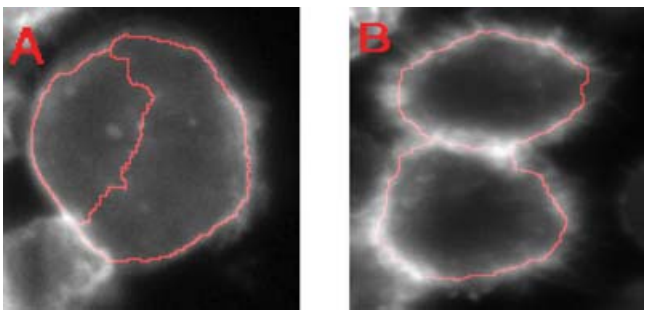

Fig. 9. Illustration of the common edges. A: Common edge between two partial cells. B: Common edge between two intact cells.

The reason we choose the minimum of the average intensities of the boundaries, $b1$ and $b2$, is to eliminate noisy pixels in the boundary that may disturb the average intensity of the boundary. If the common edge is weak, the EI score, S_{EI}^{c}, will

be high. It follows that if the common edge is true and strong, S^c_{EI} will be low.

We use three models defined above to describe specific characteristics of over-segmented cells. The MD model detects the over-segmented cells using cellular phenotypes; the CRG model uses interior gradient information, and the EI model uses boundary intensity information. Each model has distinct and important contributions to the final merging decision. In what follows, we will present a model to combine the three scores to resolve the problem of over-segmentation.

2.4.4. Merging model using Quadratic discriminant analysis. After the MD, CRG, and EI scores are calculated, one may simply apply a linear weighted combination rules to combine the three scores to determine whether or not the two cells need to be merged. However a simple linear combination does not work well because of the following two reasons. First, the values of the three scores are in different ranges, making the weights of each parameter difficult to determine. Second, the final threshold for separating the over-segmented cells is often obtained empirically, so it is not proper for the fully automated system because fixed threshold does not always work in all images. Thus, we resort to machine learning and built a non-linear classifier that uses the three scores to sort the touching cells into the 'merging' or 'not-merging' class.

The well-known statistical QDA classifier (Duda & Hart, 1973; Lindblad *et al.*, 2004), which is shown to fit our training data set well, is adopted. The training data contains 500 pairs of cells in total, in which there are 200 pairs of partial cells and 300 intact cells. The QDA classifier assumes that the data consists of several Gaussian distributions. Mathematically, given an observation set $\{x_1, x_2, \ldots, x_n \mid x_i \in R^m\}$, we assume that there are K classes, denoted as C_i, $i = 1, 2, \ldots, k$. Let π_{cj} denotes the prior probabilities of the class C_j, and $p(x_i \mid C_j)$ denotes the probability of the observation x_i for class C_j. So the posterior probability of class C_j after observing sample x_i has the following form:

$$p(C_j \mid x_i) = \frac{p(C_j, x_i)}{p(x_i)} = \propto \pi_{cj}\, p(x_i \mid C_j). \tag{6}$$

According to the Bayes rule, choosing the class $C_j$of the observation of $x_i$via maximizing the posterior probability will have the smallest expected number of errors. The QDA uses the following discriminate functions:

$$g_j(x_i) = \log(\pi_{Cj}) - \frac{1}{2}(x_i - \mu_{Cj}) \sum_{Cj}^{-1} (x_i - \mu_{Cj}) - \frac{m}{2} \log\left(\left|\sum_{Cj}\right|\right), \; j = 1, 2, \ldots, k, \tag{7}$$

where μ_{Cj}, Σ_{Cj} and π_{Cj} are estimated from the training data set.

In this study, we have two classes: merging class and not-merging class, $k = 2$. The input variable x_i takes values from the set $x_i = (S^i_{MD}, S^i_{CRG}, S^i_{EI})'$; i.g. $m = 3$. The decision of the quadratic classifier is given by:

$$y = \begin{cases} 0, & if \quad g_1(x_i) \geq g_2(x_i); \\ 1, & if \quad g_1(x_i) < g_2(x_i). \end{cases} \tag{8}$$

For one input of a pair of cells, if $y = 0$, this pair is classified into the not-merging class, if $y = 1$, this pair is classified into the merging class.

Merging procedure – feedback System. Based on QDA classification results, the over-segmented cells are detected and merged. For some cells, when we check its neighbours, we notice that there are several neighbours that satisfy the merging conditions simultaneously. In such cases, cell pairs with the largest merging probability are chosen to be merged. If none of the cell pairs belong to the merging class, the cell is tagged with a value that indicates the cell has been checked to avoid repeated check. Also, when a cell is merged with its neighbour, the new cell has two nuclei; therefore, an artificial nucleus, based on the original two nuclei, is generated (illustrated in Fig. 10). The artificial nucleus ensures that each cell had only one nucleus for the next computation of $S^c{}_{\mathrm{CRG}}$. The feedback procedure is repeated until no cell can be merged further. The procedure of the entire feedback system is described as follows:

1 Step 1. Select a cell c_0 that has not been checked before. Put all its touching neighbours $\{c_1, c_2, \ldots, c_k\}$ into one merging candidate list.
2 Step 2. Compute three scores for every pair of cells $\{c_0, c_j\}$, $j \in \{1, 2, \ldots, k\}$.
3 Step 3. Input three scores into the QDA classifier. Obtain the classes of every pair of cells $\{c_0, c_j\}$, $j \in \{1, 2, \ldots, k\}$.

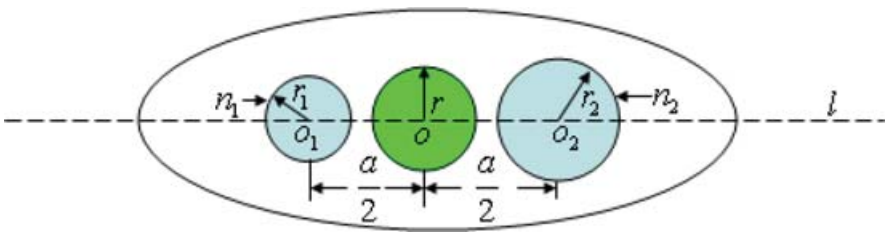

Fig. 10. Illustration of creating new nucleus, where $a = |o_1 - o_2|$ and $r = \frac{(r_1 + r_2)}{2}$.

Table 1. Confusion matrix of the QDA classifier on the extracted nuclei data set.

		Predicted	
		Dead nuclei	Live nuclei
Actual	Dead nuclei	95.83%	4.17%
	Live nuclei	4.84%	95.16%

4 Step 4. If there is no element in the merging class, tag the cell c_0. Otherwise, choose the cell pair $(c_0, c_l), l \in (1, 2, \ldots, k)$ with the largest merging probability and merge them. Generate an artificial nucleus for the new cell.
5 Step 5. Repeat steps (1) to (4) until all the cells have been tagged.

3. Results

3.1. Evaluation of the dead nuclei removal model

The first experiment is to test the model of removing the dead cell nuclei. We randomly selected six DNA channel images and applied the nuclear segmentation algorithm generating a total of 1351 nuclei with 108 dead cells nuclei. To evaluate the dead nuclei removal model, the three-fold cross-validation scheme was employed. Table 1 provides the experiment results in detail. The sensitivity level is 95.83%, and the specificity level is 95.16%.

3.2. Evaluation of the feedback system

3.2.1. Experiment on 500 pairs of selected touching cells set. To evaluate the accuracy of our feedback system, 500 pairs of cells were selected, including 200 pairs of dividing cells and 300 pairs of intact touching cells, from 200 F-actin channel images. Three-fold cross-validation method was used. For computational efficiency, five features were automatically selected for each phenotype in the MD models. Table 2 gives the number of selected features from five categories of features, and the classification accuracies on three training sets. Table 3 shows the results of our feedback system. Table 4 displays the classification results using the three individual MD score, CRT score, EI score and combination of three scores. By comparing Tables 3 and 4, it is clear that the feedback system detected dividing cells more accurately and robustly than the three scores separately.

Table 3. Confusion matrix of the QDA classifier with three scores on the extracted cells data set.

		Predicted	
		Partial cells	Intact cells
Actual	Partial cells	93.70%	6.30%
	Intact cells	5.77%	94.23%

Table 4. Classification results of the three individual scores and their combinations.

	Sensitivity	Specificity	Error rate
MD score	91.35%	60.57%	27.12%
CRI score	96.68%	90.3%	6.66%
EI score	64.2%	96.30%	16.52%
Combination of three scores	93.70%	94.23%	6.00%

3.2.2. Comparison of our feedback system with CellProfiler and manual annotation. We tested our feedback system on four Drosophila actin channel images that contained many dividing cells. Figure 11 illustrates the results of the selected cytoplasm images in which we can clearly see that several red nuclei share one common green cytoplasm in the first column. Results prior to the feedback merging step and after the feedback merging step are arranged in the middle and the right columns, respectively. It is shown that most of the over-segmented cells are detected and merged. Table 5 gives the detailed statistical result of the four cytoplasm images. We can see that 91.30% of the over-segmented cells are detected and merged via the feedback system and only 2.80% of total cells are under-segmentation.

To evaluate our feedback system, we also compared the segmentation results with that of CellProfiler, which is free software for fluorescence image analysis (http://jura.wi.mit.edu/cellprofiler/). CellProfiler resulted in considerable over-segmentation of cell cytoplasm caused by the over-segmented nuclei and the dividing cells; see Fig. 12. The method used in 'CellProfiler' is the typical seeded watershed method in which the nuclei are used as the 'seeds' to segment the cell cytoplasm. So the over-segmentation

Table 2. Three selected subsets of features and classification accuracy on the three training sets.

Features/ MD models	Wavelet	Geometry	Moment	Texture	Shape	Accuracy
Normal-MD model	1	1	1	2	0	94.75%
Spiky-MD model	1	1	0	2	1	90.33%
Ruffling-MD model	1	1	1	1	1	91.33%

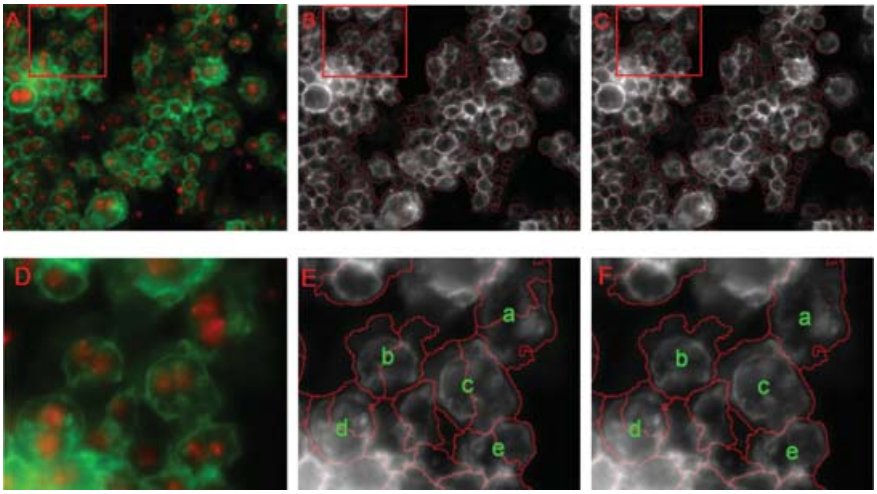

Fig. 11. Comparison of the results of segmentation before and after feedback merging. (A): the colour image with red nuclei and green cytoplasm; (B): the segmentation result before feedback merging; (C) the segmentation result after feedback merging. (D–F): the magnified images corresponding to the red box in Fig. A–C. Cells labelled a–e are over-segmented before merging (E). After feedback merging, cells a, b, c and e are appropriately merged, whereas d remains over-segmented.

problem cannot be solved because of the existence of cell division. Using the segmented nuclei directly as seeds for cell cytoplasm segmentation, as in CellProfiler, resulted in nearly 20% over-segmentation of the cells. The proposed feedback system alleviates the over-segmentation problem dramatically.

4. Conclusion and discussion

For RNAi high content screening, fully automated image analysis systems are needed urgently. Cell segmentation is the essential part of the image analysis in which over-segmentation problem is one of most challenged problem. To the best of our knowledge, there are no efficient algorithms that solve the over-segmentation problems caused by dead nuclei and dividing cells. Definiens (http://www.definiens.com/documents/) is a rather expensive commercial image processing package which use a cognitive network based algorithm to fuse the over-segmented cells. We performed evaluation using the Definiens trial version, and we found that the results generated by Definiens are rather poor. The major reasons of the poor performance may be that (1) it is difficult to choose a set of effective features in Definiens in order to describe the objects to guide the fusing process, especially in the complex RNAi cell images and (2) Definiens is not designed for RNAi cellular image segmentation.

In the present work, we have demonstrated a novel and fully automated feedback system developed for fluorescence image cellular segmentation in the context of high-throughput RNAi morphological screens of Drosophila cultured cells. We specifically focused on reducing the over-segmentation problem to obtain more accurate segmentation result. To reduce the over-segmentation problem, three scoring models were defined carefully. The advantage of the proposed method is that three novel scoring models work together can identify the over-segmented cells accurately and robustly. On the other hand, the proposed method also expands the nucleus information based cell cytoplasm segmentation methods discussed in the introduction section. The experimental results and validation of RNAi HCS images showed that the proposed feedback system reduces over-segmentation to a significant degree.

To improve the performance of the proposed method, it will be important to select a good set of training images to enhance the classifier; in this way on-line training strategy can be adopted to enhance the performance of the feedback merging system. The computational speed is also important in the RNAi high content screening project, we compared the

Table 5. Statistical result of our feedback algorithm on four actin images.

	Image I	Image II	Image III	Image IV	Total
# of over-segmented cells	27	36	40	35	138
# of correctly merged	23	34	38	31	126 (91.30%)
# of incorrectly merge	4	5	4	6	19 (2.80%)
# of total intact cells	148	178	185	168	679

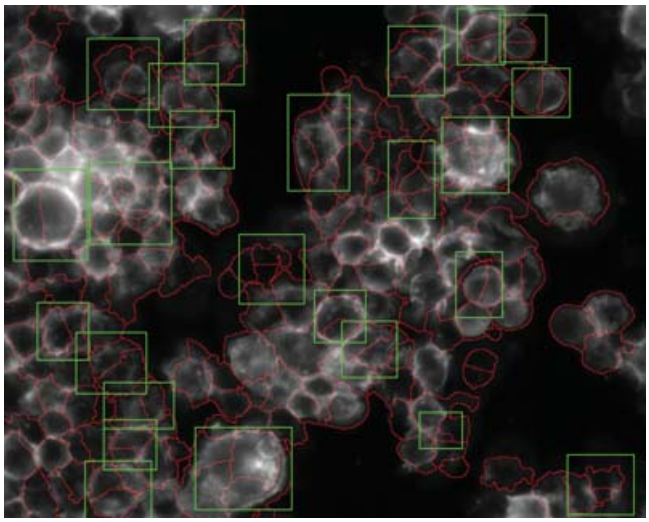

Fig. 12. The over-segmentation result of CellProfiler.

proposed method with level sets (Xiong *et al.*, 2005). Both of the methods implemented with Matlab codes, the program run on a windows XP computer (Pentium 4 2.8G). We use 10 sets of RNAi screening images (1280 × 1024 pixels) in the evaluation. The results show that the level sets method takes more than 3 days whereas our proposed method takes about 11 min, which is much faster than the level sets method. In conclusion, the proposed feedback merging system resolves the problem of over-segmentation of cell bodies in RNAi screening high-content images efficiently.

Acknowledgments

The authors would like to thank members of the Life Science Imaging Group of the Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair (HCNR), Harvard Medical School, Functional and Molecular Imaging Center, Radiology, Brigham and Women's Hospital, Harvard Medical School, and Department of Genetics and Howard Hughes Medical Institute, Harvard Medical School. The authors also would like to thank Mr. Norbert Perrimon, Ms. Pamela L. Bradley for extensive biological assistance and guidance, and Mr. Jun Wang for many good suggestions. The research is funded by HCNR to STCW.

References

Adiga, U. & Chaudhuri, B. (2001) An efficient method based on watershed amd rule-based merging for segmentation of 3-D histo-pathological images. *Pattern Recognit.* **34**, 1449–1458.

Beucher, S. (1992) The watershed transformation applied to image segmentation. *Scanning Microsc. Int.* **6**, 299–314.

Boutros, M., Kiger, A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S., Paro, R. & Perrimon, N. (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science* **303**, 832–835.

Chan, T. & Vese, L. (2001) Active contours without edges. *IEEE Trans. Image Process.* **10**, 266–277.

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

Cohen, A., Daubechies, I. & Feauveau, J. (1992) Biorthogonal bases of compactly supported wavelets. *Commun. Pure App. Math.* **45**, 485–560.

Duda, R. & Hart, P. (1973) *Pattern classification and scene analysis*. John Wiley & Sons, New York, USA.

Haralick, R. (1979) Statistical and structural approaches to texture. *Proc. IEEE* **67**, 786–804.

Jain, A. & Zongker, D. (1997) Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Patt. Anal. Mach. Intell.* **19**, 153–158.

Jones, T., Carpener, A. & Golland, P. (2005) Voronoi-based segmentation of cells on image manifolds, *Lecture Notes in Computer Science*, 535–543.

Kapur, R. (2001) High content screening and the CellChip-TM system: living cells as beacons for drugs and toxins. *Eur. Cells Mater.* **2**, 7.

Kass, M., Witkin, A. & Terzopoulos, D. (1987) Snake: Active Contour Models. *Int. J. Comp. Vis.* **1**, 321–331.

Kiger, A., Baum, B., Jones, S., Jones, M., Coulson, A., Echeverri, C. & Perrimon, N. (2003) A functional genomic analysis of cell morphology using RNA interference. *J. Biol.* **2**(4), 27.

Klemencic, A., Koveacic, S. & Pernus, F. (1998) Automated segmentation of muscle fiber images using active contour models. *Cytometry A* **32**, 317–326.

Lin, G., Adiga, U., Olson, K., Guzowski, J., Barnes, C. & Roysam, B. (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.

Lin, G., Chawla, M., Olson, K., Guzowski, J., Barnes, C. & Roysam, B. (2005) Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei. *Cytometry A* **63**, 20–33.

Lindblad, J., Wahlby, C., Bengtsson, E. & Zaltsman, A. (2004) Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation. *Cytometry A* **57**, 22–33.

Malladi, R., Sethian, J. & Vemuri, B. (1995) Shape modeling with front propagation – a level set approach. *IEEE Trans. Patt. Anal. Mach. Intell.* **210**, 404–415.

Manjunath, B. & Ma, W. (1996) Texture features for browsing and retrieval of image data. *IEEE Trans. Patt. Anal. Mach. Intell.* **18**, 837–842.

Morelock, M., Hunter, E., Moran, T., Heynen, S., Laris, C., Thieleking, M., Akong, M., Mikic, I., Callaway, S., DeLeon, R., Goodacre, A., Zacharias, D. & Price, J. (2005) Statistics of assay validation in high throughput cell imaging of nuclear Factor κB nuclear translocation. *ASSAY Drug Develop. Technol.* **3**, 483–499.

Otsu, N. (1978) A threshold selection method from gray level histgram. *IEEE Trans. Syst., Man, and Cybern.* **8**, 62–66.

Perlman, Z., Slack, M., Feng, Y., Mitchison, T., Wu, L. & Altschuler, S. (2004) Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198.

Pham, T., Crane, D., Tran, T. & Nguyen, T. (2004) Extraction of fluorescent cell puncta by adaptive fuzzy segmentation. *Bioinformatics* **20**, 2189–2196.

Pudil, P., Novovičová, J. & Kittler, J. (1994) Floating search methods in feature selection. *Patt. Recognit. Lett.* **15**, 1119–1125.

Sahoo, P., Soltani, S., Wong, A. & Chen, Y. (1988) A survey of Thresholding Techniques. *Computer Vision Graphics Image Processing* **41**, 233–260.

Siedlecki, W. & Sklansky, J. (1989) A note on genetic algorithms for large-scale feature selection. *Patt. Recognit. Lett.* **10**, 335–347.

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

Wahlby, C., Lindblad, J., Vondrus, M., Bengtsson, E. & Bjorkesten, L. (2002) Algorithms for cytoplasm segmentation of fluorescence labelled cells. *Anal. Cell Pathol.* **24**, 101–111.

Wang, J., Zhou, X., Li, F. & Wong, S. (2006) *Classify Cellular Phenotype in High-Throughput Fluorescence Microcopy Images for RNAi Genome-Wide Screening*. IEEE/NLM Life Science Systems & Applications Workshop. Bethesda, MD, July 2006, 1–2.

Xiong, G., Zhou, X., Ji, L., Bradley, P., Perrimon, N. & Wong, S. (2005) Automated segmentation of Drosophila RNAi fluorescence cellular images using deformable models. *Circuits Syst., IEEE* **53**, 2415–2424.

Zernike, F. (1934) Beugunstheorie des schneidencerfarhens undseiner verbesserten form, der phasenkontrastmethode. *Physica* **1**, 689–704.

Zhou, X., Liu, K., Bradley, P., Perrimon, N. & Wong, S. (2005) Towards automated cellular image segmentation for RNAi genome-wide screening. *Lecture Notes in Computer Science (MICCAI 2005) Vol 3749*, Spring-Verlag, Berlin, 885–892.

Zhou, X. & Wong, S. (2006a) High content cellular imaging for drug development. *Signal Process. Mag., IEEE* **23**, 170–174.

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

