On the Accurate Counting of Tumor Cells

Bin Fang[a] Wynne Hsu[a,b] Mong Li Lee[b]

[a]Singapore-MIT Alliance
National University of Singapore

[b]Department of Computer Science, School of Computing
National University of Singapore

Abstract

Quantitative analysis of tumor cells is fundamental to pathological studies. Current practices are mostly manual, time-consuming and tedious, yielding subjective and imprecise results. To advance the understanding of tumor cells behaviour, it is critical to have an objective way to count the number of tumor cells. In addition, these counts must be reproducible and independent of the person performing the count. To achieve this, we propose a two-stage tumor cell identification strategy. In the first stage, potential tumor cells are segmented automatically using local adaptive thresholding and dynamic water immersion techniques. Unfortunately, due to histological noise in the images, a large number of false identifications have been made. To improve the accuracy of the identified tumor cells, a second stage of feature rules mining has been initiated. Experiment results show that image processing techniques alone are unable to give accurate results for tumor cell counting. However, with the use of features rules, we are able to achieve an identification accuracy of 94.3%.

1. Introduction

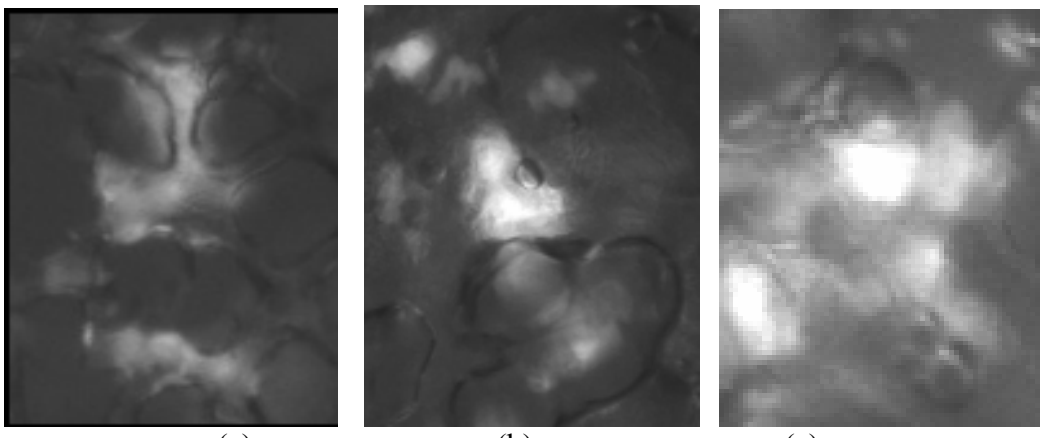
The mechanism of tumor cell metastasis has been the subject of research for many years in pathology [1,2]. Tumor cells first migrate from the primary tumor, penetrate into the circulation, and eventually colonize distant sites. Knowledge regarding the dissemination of tumor cells is very important in clinical studies of pathology. The quantitative analysis of tumor cells in the filed of pathology forms the basis for characterizing the dissemination activity of tumor cells. To perform quantitative analysis of tumor cells, the original tumor cells are first stained with special materials such as GFP – Green Fluorescent Protein. These stained cells are then being introduced into the experiment animals. After a few days, tissue sections containing the stained tumor cells are harvested from certain body organ such as lungs. A trained medical professional will then manually count the number of tumor cells expressing GFP in the tissue section under a converted fluorescence microscope focused onto the tissue section or on a monitor screen displaying fluorescence cell images that are digitally

captured from the tissue sections. This process is laborious and tedious, yielding subjective and imprecise results. Hence, there is an increasing demand for an automated system that can analyze the digitized histological images and identify tumor cells accurately.

The design of such an automatic segmentation system for biomedical image analysis in pathology has been a major research objective for many years. Automatic segmentation methods are generally based on local image information such as pixel intensity, discontinuity of intensity, histogram or clusters. Traditionally, segmentation techniques are categorized into two classes: those that employed region-finding algorithms versus those that employed contour-detection algorithms. Region-finding algorithms [3, 4, 5, 6] involve partitioning the grey level histogram in such a way that the appropriate thresholds for segmentation can be easily determined. Contour-detection algorithms rely on the discontinuity of the image intensities or texture at the boundary of objects [7, 8]. Contour-finding algorithms generally have poor performance in noisy images as the presence of noise tends to lead to broken and disjoined edges which require additional efforts to analyze these broken edges before boundaries information can be extracted. Although region-finding algorithms are less sensitive to image noise, they are usually computationally more expensive.

There have been a lot of cell segmentation methods that employ image processing techniques to deal with case-specific problems reported in the literature [9-16]. Jeacocke *et al*. [9] used a multi-resolution method which contains quadtree smoothing, lowest level classification and boundary re-estimation by water immersion. Chen *et al*. [10] used spatial adaptive filter, watershed and refining of the labeled image. Wu *et al.* [11] presented a parametric fitting algorithm for segmentation of single cervical and breast cell images from cytology smears by incorporation *a prior* knowledge of the objects to be identified as elliptical curves in the paper. However, in most situations, *a prior* information is not be available. Thresholding methods based on different kinds of statistical information such as grey-level histogram and various mathematics morphological operations were also used for segmentation. Kovalev *et al*. [12] used intensity histogram of G color plane and the balance between G and B color intensity for color images. Awasthi *et al*. [13] used a combination of multiple thresholding, dilation morphology operation and region growing methods to perform cell segmentation. Berns *et al*. [14] used a combination of median filter, local histogram, and morphology filter with watershed method to achieve the same goal of cell segmentation. C. G. Loukas *et al*. [15] generated Principal Component (PC) and processed histogram of the PC

with greatest contrast by an appropriate threshold. The region of interests (ROIs) were detected by constructing a distance map for every particle. Anoraganingrum [16] used median filter and mathematical morphology operation for edge detection based cell segmentation. However, these methods are unable to effectively segment the ROIs with histological noise and non-uniform background. Wu *et al.* [17] proposed a two-step segmentation strategy to deal with the problem of uneven illumination in images. First an approximate region containing the single cell and part of the background near the cell is detected by applying a global threshold to the local variation of intensity. Second the cell is segmented from the background within such approximate region by Otsu's method. The drawback of the method is that the computation is very expensive and clumped cell clusters are not taken into account. There are few works reported on extracting individual cells from segmented ROIs corresponding to clumped cell clusters. Mathematical morphology operations and watershed techniques lead to over-segmentation and result irregular contours of cells. C. G. Loukas *et al.* [15] used either a standard morphological operation called Skeleton by Influence Zone or a proposed heuristic processing of a distance map resulted from former process step to detect individual cells. A.K. Jain *et al.* [18] employed a hierarchical clustering algorithm to group together those boundary points of a cell clump that belong to the same globally convex sections of the boundary. When the segmented ROIs do not provide sufficient shape complexity information for convex or concave analysis and cell boundaries inside the cell clusters are not sharp enough to be readily extracted, these methods generally seem unable to successfully detect individual cells.

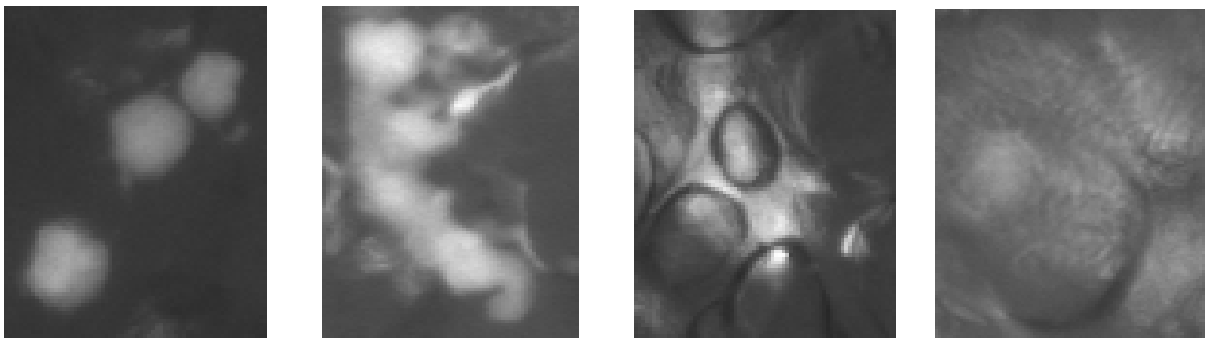


(a) (b) (c)

Figure 1. Portions of a tissue-section histological image.

Figure 1 shows the highly complex characteristics of some typical fluorescence cell images of tissue sections in the present study. Most of these images exhibit non-uniform background illumination. In particular, the tumor cell boundaries are not sufficiently sharp to be readily extracted and it can be observed that many tumor cells, in fact, join with each other to form clumped cell clusters throughout the images. We make the following observations:

1) High intensity pixels in the form of white patches scattering in the image may have a number of interpretations: they could be the individual tumor cells (Figure 2a), or they could be clumped cell clusters where the tumor cells have grown to form a colony (Figure 2b), or they could be histological noise such as reflection of light on the spherical surface of normal cells (Figure 2c), or they could be pseudo-pods of tumor cells (Figure 2b).
2) The non-uniform background illumination may result in the intensity of the tumor cells being darker than the background intensity at distant location of the image (Figure 2d).
3) The presence of cell pseudo-pods complicates the problem further as cell boundaries are typically not sharp enough to be readily extracted. In addition, there are many situations where the cells are joined or located so close to one another that highly irregular shapes are formed (Figure 2b).
4) The intensity of a tumor cell usually peaks at the center of the cell. However, when two cells are touching an intensity valley will be formed between the two peaks (Figure 2b).



(a) (b) (c) (d)

Figure 2. (a) Individual tumor cells, (b) clumped tumor cell clusters and pseudo-pods of tumor cells, (c) histological noise of light reflection, (d) 'brighter' background.

Due to the complex characteristics of the images in our study, it is evident that no existing case-specific methods for cell segmentation can produce satisfactory results. Moreover, our studies indicate that image processing techniques alone are unable to give accurate results for

tumor cell counting. Instead, we propose a two-stage tumor cell identification strategy. First, we propose a local adaptive thresholding approach to segment regions of interest from the background by minimizing the negative effects of distant background noise. We also propose a dynamic water immersion algorithm to detect individual tumor cells in clumped tumor clusters so that these individual tumor cells can be extracted as objects of interest. Investigations show that our proposed approach is able to effectively extract reliable circle-like contours of the tumor cells, as well as contours that correspond to histological noise such as white light reflection areas and the extended pseudo-pods of tumor cells. In other words, pure image processing techniques alone are unable to provide a high accuracy for the identification of tumor cells. To overcome this problem, we propose the use of feature rules to help distinguish tumor cells from histological noise.

Applying feature rules to aid in pattern identification is rapidly gaining popularity in the field of biomedical engineering. Thiran *et al.* [19] employed the evaluation of an overall score for a digital image based on extracted features measured on individual cells to decide whether the tissue image is cancerous or not. However, they do not perform classification on individual cells using the extracted features. Chan *et al.* [20] compared a variety of machine classifiers built from the STATPAC indexes traditionally used for glaucoma diagnosis on SAP data. The SAP data are scalar values and are collected from existing database. Properties, advantages and disadvantages of generative and discriminative machine classifiers are applied to the SAP data have been compared. Nattkemper *et al.* [21] used an artificial neural net of local linear map for classifying fluorescent lymphocytes in tissue sections. The system is semi-automatic that it acquires visual knowledge from a set of training cell-image patches selected by a user. Antonie *et al.* [22] investigated the use of neural network and association rule mining for classifying digital mammograms that were segmented into two categories: normal and abnormal images. Hsu *et al.* [23] combined 12 image attributes extracted for each individual vessel segment and fed them into an association based data mining classification tool (CBA) to classify the input vessel segments as normal or abnormal in the application of an Integrated Retinal Information system. Vessel segments were obtained by proposed case-specific image processing techniques. L. Zheng *et al.* [24] presented several artificial intelligence techniques to detect candidate regions in mammogram. A tree-type classification strategy was then applied at the end to determine whether a given region was suspicious for cancer. However, as far as we know, there has

been no application of data mining techniques to the problem of tumor cell identification for fluorescence cell images.

In this paper, relevant meaningful features are extracted from objects of interest and three base classifiers are built to generate features rules that will differentiate tumor cells from histological noise. Further, since different classifiers may potentially offer complementary information about the patterns to be classified, they could be integrated to improve the overall performance of the identification system [25, 26]. In this study, we unify three base classifiers (decision-tree based, association-rules based, and probability-based) into a meta-classifier using the majority vote strategy. Experiment results indicated that the second phase of using features rules to further enhance the accuracy of tumor cells identification is effective. We are able to achieve an accuracy of 94.3% in the identification of tumor cells.

2. Overview

Figure 3 shows an overview of our proposed approach. First, we apply a local adaptive thresholding method to segment the white patches as objects of interest. Second, we use a dynamic water immersion approach to locate the individual cells within clumped cell clusters. These individual cells are added to the set of objects of interest for further investigation. For each object of interest, we extract several discriminative features to accurately describe the characteristics of the extracted regions. Next, based on the extracted features, three classifiers are built to differentiate the tumor cells from noise. They are the Naïve Bayes classifier [27], C4.5 classifier [28] and CBA classifier [29]. Finally, a meta classifier with majority voting strategy is also implemented on top of the three base classifiers to further improve the accuracy for tumor cells identification.

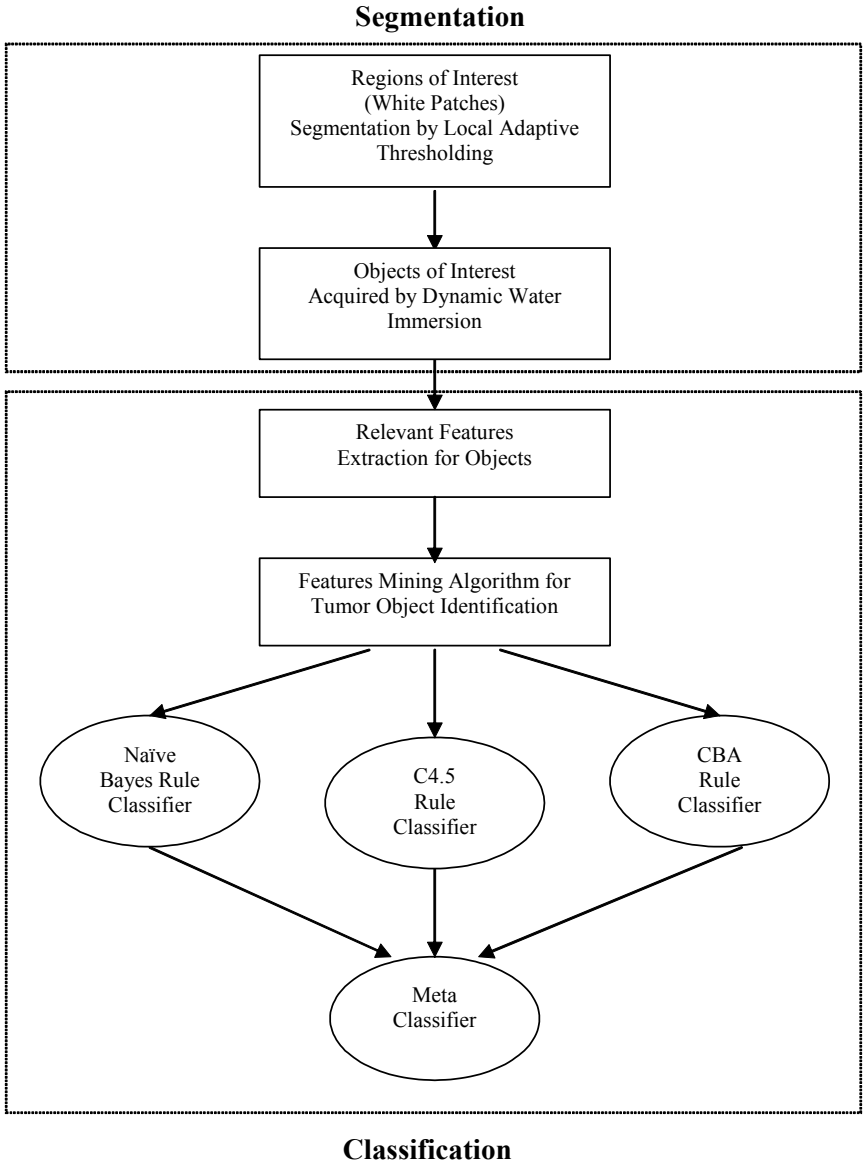


Figure 3. Overview of tumor cell identification process.

3. Objects of interest extraction

The most crucial step in this tumor cell identification system lies in the effective extraction of the objects of interest which forms the basis for the next phase of feature-based classification. To accurately extract all objects of interest without missing any potential tumor cells, we adopt a two step process. The first step segments meaningful regions from the highly non-uniform background. The emphasis in this step is to minimize the negative effects of distant background noise. The exact shape and size of the segmented regions are not important. The

second step aims to separate the touching or closely located tumor cells out so as to obtain the individual objects of interest related to potential tumor cells.

3.1 Local adaptive thresholding

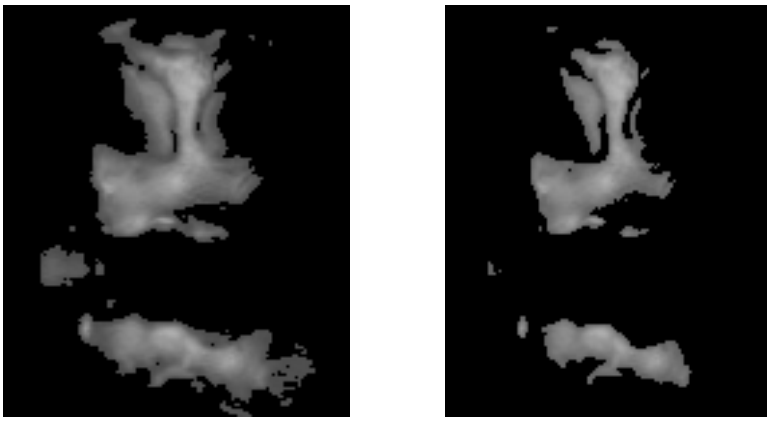
Segmentation of tissue images remains an open problem as described in Section 1. A popular implementation of region-finding segmentation methods is thresholding which can be regarded as pixel classification [30]. Thresholding methods can be classified as either global or local. Global thresholding techniques use a single threshold level for the entire image [3, 4] while local methods [31, 32], similar in principle to contour-finding algorithms, develop a threshold surface which allows a different threshold value to be applied to each pixel. Local thresholding methods seem to produce unconnected boundaries of objects and various linking operations are needed for final object segmentation. In our case, the global thresholding method is unable to effectively segment all white patches that are scattering over the image. This is because that background is non-uniformly illuminated causing white patches at one location in the image to be 'darker' than the background at other locations. Applying a global thresholding across the entire image would result in 'brighter' background regions being misclassified as cells and 'darker' cell regions being misclassified as background.

Figure 4(a) shows segmented result of a local portion of the entire image by applying the global thresholding method proposed by Otsu [4]. Clearly, a large number of unexpected background noises remain in the segmentation result for the image given. To deal with this problem, we propose the use of a local adaptive thresholding scheme based on the analysis of pixel intensity distribution in a specific sub-image. First, the image is divided into N×N sub-images of equal size. Then, the histogram of each sub-image is computed and the local threshold is determined. According to the mean and variance of intensity histogram measured on the sub-image, the threshold is set as follows:

$$TH = mean + \alpha \times std \qquad (1)$$

where *TH* is the adaptive threshold, *mean* and *std* represents the mean and standard deviation of the intensity distribution of all pixels in the sub-image respectively. α is a constant. Figure 4(b) shows the result of applying local adaptive thresholding method to segment regions of interest from background with α = 1.0. The segmented white patches cover all potential tumor cells as identified by a medical professional with only a small percentage of

background noise. Hence, the local adaptive thresholding method is able to provide adequate segmentation results of white patches with histological meaning for further process.



(a) (b)

Figure 4. Segmentation results of regions of interest applied to the image in Figure 1(a). (a) by Otsu's global thresholding method. (b) by the proposed local adaptive thresholding method.

3.2 Dynamic water immersion

Having obtained the segmented regions of interest, our next task is to separate the clumped cell clusters into individual cell objects so that meaningful features for each object can be extracted for further mining process. This problem is made complicated by the fact that the boundaries of the cells are fuzzy and cannot be readily extracted. Furthermore, the shapes of cells are deformed due to the presence of cell pseudo-pods. Edge detection techniques such as Laplace of Gaussian (LoG) method perform poorly as it tends to lead to broken and disjoined edges which require additional efforts to analyze the output of an edge detection image before boundaries information can be extracted.

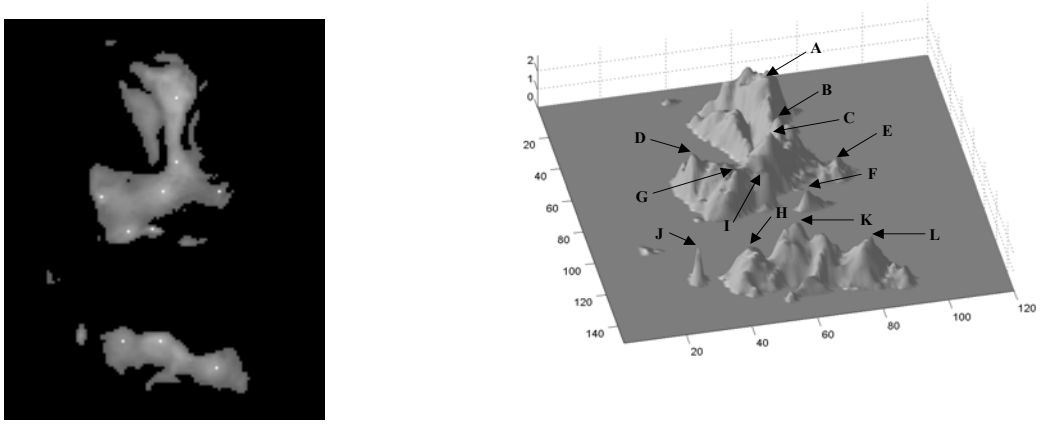
Watershed or water immersion algorithm is considered to be a powerful technique for touching object contour detection [33, 34]. Water immersion algorithm works by grouping pixels with similar gradient information. Direct application of water immersion method to the digitized histological images typically produces over-segmentation of the individual cells. Instead, we propose a dynamic water immersion algorithm to cope with the situation.

First, a sliding N×N window is used to locate the local peaks with maximum intensity in the regions of interest. For each segmented white patches, we place the center of the window over each pixel in the white patches. If the intensity of the center pixel is the highest

with respect to all the other pixels in the window, we say that the center pixel is a local peak; otherwise, the window will move to be centered at another pixel to continue the search for all local peaks. At the end of this phase, all the peaks are marked and they will be treated as the starting seeds for water immersion method.

One advantage of using the sliding window approach is that with the appropriate window size, it is possible to eliminate a large amount of peaks that correspond to the light reflection regions thus removing false detection. This is because the intensity levels of the peaks corresponding to the light reflection areas are generally lower than that of potential tumor cells or even the extended pseudo-pods of cells. Given that the distances between the peaks of the light reflection patches and the nearest peaks of the neighboring tumor cells are generally less than that between two touching tumor cells, it is possible to set the window size in such a way that these false peaks are 'absorbed' by the neighboring tumor cell peaks while the true peaks corresponding to the touching cells are not affected. Another advantage of applying the sliding window with suitable size is to eliminate peaks within smaller isolate regions. This is done by setting an area threshold based on the window size (one fifth in the experiments). Any regions with area smaller than the threshold will not be considered for peaks detection. Figure 5(a) shows the result of choosing a suitable window size on the detection of peaks. Figure 5(b) illustrates with 3D representation that points *F* and *J* are in small regions and they are not marked out as peaks for water immersion algorithm.

Having identified the intensity peaks, water immersion process starts from the detected peaks denoted as seeds and progressively immerses its neighboring pixels. The neighboring pixels are defined to be the 8-direction neighbors. These neighbors are placed in a growing queue structure sorted in descending order of the intensity level of the pixels. The lowest intensity pixel in the growing queue will be 'immersed' first and it is marked as belonging to the same object label as the current seed. The marked pixel is then removed from the growing queue. All neighboring pixels whose intensity level is lower than the marked pixel are added to the growing queue. This progressive immersion process continues until the growing queue is empty.



(a) (b)

Figure 5. (a) Results of intensity peaks detection applied to the image in Figure 1(a). Local intensity peaks are marked by bright white dots. (b) 3D representation of the same segmented regions of interest illustrates local intensity peaks as seeds for water immersion algorithm. Points F and J are in small isolate regions and not determined as peaks.

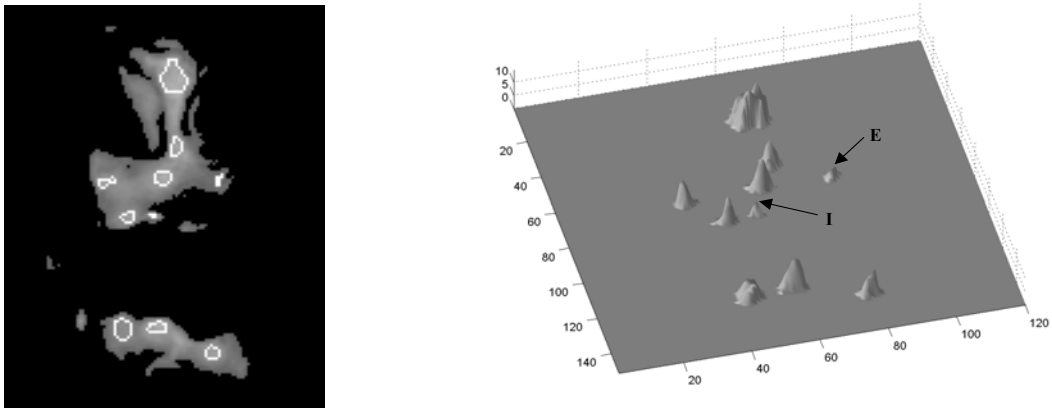
The direct application of the water immersion technique has the tendency of over-immersion that leads to incorrect and deformed contours of the tumor cells. The tumor cells should be circular but instead, the resulted cell shapes are quite irregular. This can seriously undermine the accuracy of the classifiers which aim to distinguish tumor cell objects from noise objects. To overcome this tendency of over-immersion, we apply an additional stop criterion. Besides ignoring all neighboring pixels with intensity level lower than the last placed pixel, we also ignore all those pixels whose intensity level is too low as compared to the seed pixel. To implement this, we use a dynamically set seed-to-pixel contrast threshold. This threshold is larger for the 'brighter' seed and smaller for the 'darker' seed. This is because from a priori knowledge, we know that the variation in intensity level of pixels neighboring to a 'brighter' peak is larger than that to a 'darker' peak in which case the contour related to the 'brighter' peaks should be longer. The seed-to-pixel contrast threshold is determined directly relating to the intensity quantity of the seed involved using the equation as follows:

$$Con\_th = A \times \frac{I_{\max}}{255} \times e^{\frac{I_{\max}}{255}} \qquad (2)$$

where *Con_th* is the seed-to-pixel contrast threshold, A is a constant determined by analyzing the intensity level variation of the tumor cell objects , and I_{max} is intensity level of the seed. The pseudo-codes for the dynamic water immersion algorithm are given as follows.

Algorithm DynamicWaterImmersion

```
Count_Q = 1
FOR each seed of local intensity maxima s
        I_s = intensity value of the seed
        Con_th = seed-to-pixel contrast threshold
        FOR each 8-directional neighbor of the seed s, P_i
                I_p = intensity value of P_i
                IF  (I_p > I_s )  THEN
                        P_i → Queue; Count_Q=Count_Q + 1
                ENDIF
        ENDFOR
        LOOP
                Sort elements in Queue in descent order
                IF ( L_i is at last position in Queue) THEN
                        I_l = intensity value of L_i
                        FOR each 8-directional neighbor of L_i not in the Queue, R_i
                                I_r = intensity value of R_i
                                IF  ( I_r > I_l  .and. abs(I_r - I_s) < Con_th)  THEN
                                        R_i → Queue; Count_Q=Count_Q + 1
                                ENDIF
                        ENDFOR
                        Li is removed from the Queue and marked the same object label as
                        the seed.
                        Count_Q = Count_Q – 1
                ENDIF
                IF ( Count_Q = 0 ) THEN exit LOOP
        ENDLOOP
ENDFOR
```



(a) (b)

Figure 6. (a) Illustration of extracted objects of interest marked by continuous white line applied to the image in Figure 1(a). (b) 3D representation of the same objects of interest illustrates that objects related to noise (E,I) have distinct characteristics from tumor cells such as smaller object volume.

Figure 6(a) shows that the proposed algorithm is able to effectively extracted reliable contours of the tumor cells, and is able to separate those touching tumor cells. At the same time, a number of light reflection areas have been eliminated. However, the algorithm still extracts false objects corresponding to the pseudo-pods of tumor cells as shown in Figure 6(b). With the existance of histological noise objects, the counting of tumor cells is not accurate. In the next section, we discuss how features mining algorithm can be used to increase the identification accuracy by automatically classifying these objects of interest into either tumor cells or non-tumor cells (histological noise).

4. Tumor cell identification by features mining

4.1 Feature extraction for tumor cell identification

At this point, we have extracted individual objects of interest from tissue section images. For each extracted object, we need to extract relevant information for features mining to take place. A priori knowledge of the tumor cells' characteristics indicates that touching cells and cell pseudo-pods resulted in irregularly-shaped tumor cells. Fortunately, with the proposed dynamic water immersion method, we are able to extract individual cells that conform to the circular appearance as shown in Figure 6(a). Moreover, the extracted objects that are due to histological noise usually bear different appearances from tumor cell objects in the sub-images. Compared to tumor cells, objects due to histological noise can be characterized by lower intensity peaks (maximum intensity value of the object), smaller area and width measurement, smaller 3D volume measure and more elongated appearance complexity (Figure 6(b)). Hence, the following relevant features is able to provide sufficient discriminative power to classify objects into tumor cells and histological noises: (i) maximum intensity value of the object of interest (seed intensity value), (ii) power of the object of interest, (iii) elongation of the object of interest, (iv) area of the object of interest, and (v) width of the object of interest.

The power of an object of interest is defined as follows. First, the minimum intensity value I_m of the object is searched and stored. Intensity value of each element in the object is subtracted by I_m. The power is the sum of the modified intensity values of all elements in the object.

$$P = \sum_{k=1}^{N} (I_k - I_m) \quad (3)$$

where P is the volume of an object of interest, N is the number of pixels enclosed in the object, I_k is the intensity value of object pixels and I_m is the minimum intensity value in the object.

An important morphological feature we employ to characterize the appearance of an object is the elongation measure. The elongation of an object is defined as the ratio of the width of the minor axis to the length of the major axis. This ratio is computed as the minor axis width distance divided by the major axis length distance, giving a value between 0 and 1. If the ratio is equal to 1, the object is roughly a square or is circular in shaped. As the ratio decreases from 1, the object becomes more elongated. Major axis is the longest line that can be drawn through the object. The two end points of the major axis are found by selecting the pairs of boundary pixels with the maximum distance between them. This maximum distance is also known as the major axis length. Similarly, the minor axis is defined as the line that it is perpendicular with respect to the major axis and the length between the two end points of the line intersected with the object is the longest which is called the width of the minor axis. The ratio, *Elong*, is a measure of the degree of elongation of an object.

$$Elong = \frac{L_{MAJOR}}{L_{MINOR}} \quad (4)$$

where L_{MAJOR} is the major axis length of the object and L_{MINOR} is the minor axis length of the object.

In addition, two additional features are extracted: area of the objects of interest (total number of pixels within the object, up to and including the boundary pixels), and width of the object (length of minor axis of an object of interest). These measures are useful in characterizing the appearance of connected objects. To ensure our objects of interest are connected set of pixels with no holes inside, we perform a closing morphological operation to fill all the holes that may possibly exist in the objects of interest. This ensures that each object of interest has a single border.

4.2 Features mining by base and meta classifiers

With the extracted features of each objects of interest, a number of base classifiers are built to help classify these objects into either tumour cells or non-tumour cells (histological noise).

Here, we have selected the three most commonly used data mining techniques to build our team of base classifiers. The first base classifier is the Naïve Bayes classifier [27] which applies statistical method for pattern classification. Assuming the sample distributions of all classes are multivariate Gaussian distribution:

$$f_i(\mathbf{x} \mid \omega_i) = (2\pi)^{-p/2} \mid \Sigma_i \mid^{-1/2} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] \qquad (5)$$

where $1 \le i \le L$, i is the label of the different classes, L is the number of total classes which is equal to two in our study: class of tumor cells and class of non-tumor cells.

The mean vector and the covariance matrix for each class could be computed from the training data by the maximum likelihood estimates. If there are N training samples, the statistics of the extracted features can be represented of the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of these vectors, where

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{F}_{\kappa} \qquad (6)$$

$$\Sigma = \frac{1}{N}\sum_{k=1}^{N} (\mathbf{F}_{\mathbf{k}} - \boldsymbol{\mu})(\mathbf{F}_{\mathbf{k}} - \boldsymbol{\mu})^T \qquad (7)$$

where k = 1, 2, …, N.

In order to save computation load and evaluate reliable statistics, features are usually assumed to be independent which means the diagonal elements of the covariance matrix are zeros leading to a Naïve Bayes classifier. In this case, only the variances each feature for the different classes are computed.

When a test data comes in, the classification rule will assign the test sample to the class with the highest posterior probability which can be written as follows.

$$p(\omega_i \mid \mathbf{x}) = \frac{f_i(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} \qquad (8)$$

where $1 \le i \le L$, i is the label of the different classes, L is the total number of classes, $P(\omega_i)$ is the class frequency and $p(\mathbf{x})$ is a common factor for all classes. Furthermore, the classification rule can be simplified by using the discriminant function which is defined below.

$$d_i(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln(\mid \boldsymbol{\Sigma}_i \mid) - 2\ln(P(\omega_i)) \qquad (9)$$

The test sample is classified to the class which results the smallest value of the discriminant function.

However, the true class distributions under consideration may not be Gaussian. In such cases, classification techniques that do not require such assumption may result in better performance. Therefore, we use the second base classifier based on the decision tree techniques that do not make the Gaussian distribution assumption. The most widely cited decision tree classification tool, C4.5 [28] is used to build this base classifier. C4.5 is based on ID3. The basic ideas behind ID3 are: 1) each node of the decision tree corresponds to a feature and each edge to a possible value of that feature. A leaf of the tree specifies the expected value of the feature for the samples described by the path from the root to that leaf. 2) Each node in the decision tree should be associated with the feature which is *most informative* among the features not yet considered in the path from the root. 3) *Entropy* is used to measure how informative is a node. C4.5 is an extension of ID3 that accounts for missing values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. For further details, refer to [28].

Our third and final base classifier is based on association rules mining technique. This classifier is built using the Classification-Based on Association (CBA) tool that was first proposed in 1998 [29]. The CBA tool integrates classification rule mining technique and association rule mining techniques focusing on mining a special subset of association rules. Experimental results show the CBA based classifier is usually more accurate than C4.5. Check [29] for more information.

In this study, we adopt 10 fold cross-validation testing strategy to train classifiers and evaluate classification accuracy. For each round of testing, one tenth of collected samples are left as testing samples while the remaining samples are used to model the classifier. Classification measures such as false-positive and false-negative are computed accordingly. After 10 rounds of testing, the classification measures are averaged and this average value is reported as the final classification accuracy. From past experiments, it is known that different classifiers have different abilities in classifying the different classes [14, 15]. To take full advantage of their different strengths and to examine the further improvement of predictive accuracy, we propose the use of majority voting strategy to integrate the base classifiers into a meta-classifier. In the majority vote strategy, the meta-classifier will output the majority class among all the base classifiers as the final class. For example, in our problem, we have two classes: tumor versus non-tumor. Once test data comes in, if the Naïve Bayes classifier

identifies it as class tumor, but both the C4.5 and CBA classifiers classify it as class non-tumor, then the majority-vote meta-classifier will assign the test data to class non-tumor.

5. Experimental results

The images used in our experiments are derived from the tissue sections containing tumor cells of experimental animals' lungs which had been stained using GFP - green fluorescent protein. Female mice, 4–8 weeks of age, were given injections with a single cell suspension of 2 x 10^6 cells in 100 μl into the tail vein. After 24 or 48 hours, the mice were sacrificed and the lungs were frozen in histological preparation. Ten-micrometer frozen sections were made. The section slides were scanned using a digital micrometer (Microcode II; Boeckeler Instruments) to ensure that all areas were counted only once. Green fluorescent cells were confirmed by overlay with a DAPI-stained nucleus.

The tissue sections were observed by a Leica inverted fluorescence microscope. A Hamamatsu Orca digital camera was connected to the microscope and linked to a Mac G4. A ×40 objective was used during acquisition. Fluorescence cell images were captured directly using the digital camera as 8-bit gray-level 1024×1022 TIFF files and stored on hard disk of Mac G4 computer. The digitized histological images were characterized by high intensity pixels in the form of white patches corresponding to potential tumor cells or clumped cell clusters expressing GFP. The proposed methods were evaluated on a database of 40 tissue-section histological images captured under same environment such as subjective magnification, exposure and data format, etc. The resolution of these images is 1024×1022 in 8-bit grey level TIFF format.

During the image processing stage of local adaptive thresholding, the images are divided into 3×3 sub-images. After performing some initial studies, we set α in Eqn (1) to 1.00. This value is good enough to retain as many white patches as possible while removing most of the background noise. For the dynamic water immersion algorithm, a window size of 7x7 pixels is used to locate the local intensity peaks. The *Con_A* in Eqn (2) was determined by initial investigation to be 42 for extracting contours of objects of interest. Figure 4(b) shows an example of the segmented white patches by the local adaptive thresholding method. Local peaks are detected and marked by bright white dots as shown in Figure 5(a). Contours of objects of interest extracted by using the dynamic water immersion algorithm are highlighted by continuous white lines as shown in Figure 6(a).

In our experiments, there are 9581 objects of interest extracted by the improved image processing techniques for the 40 fluorescence cell images while a medical professional had carefully labeled 1974 tumor cells. All of the tumor cells labeled by the expert are automatically detected and included in the objects of interest. These extracted objects of interest not only correspond to tumor cells but histological noise objects as well due to the complexity nature of the tissue section images processed. The histological noise objects are pseudo-pods of tumor cells and white light reflection areas which are not excluded by the process of water immersion method. Hence, using the improved image processing techniques alone, we are only able to achieve an accuracy of 20.6%, which is absolutely unaccepted. To improve the identification accuracy, features rules are examined by building a number of classifiers to help identify the tumor cells.

To evaluate the effectiveness of class assignments by individual classifiers, the false-positive and the false-negative error rates are often used as system performance measurement. False-positive error rate refers to misclassification of tumour cell as non-tumour cell and false negative error rate relates to misclassification of non-tumour cell as tumour cell. The average error rate is defined as the ratio of all misclassified assignments to all test samples. However, if class distribution is seriously skewed as in the present study, the frequency of objects of interest due to histological noise (class 0) is 7607 (79.4% of total samples), additional measures such as recall, precision and F_1 measures [35] should also be employed to evaluate the classifier performance. Recall is defined to be the ratio of correct assignments of a class to the total number of points in that class. Precision is the ratio of correct assignments of a class to the total number of the system's assignments to that class. The F_1 measure combines recall (r) and precision (p) with an equal weight in the following form:

$$F_1(r, p) = \frac{2rp}{r + p} \qquad (10)$$

Although application of the improved image processing techniques resulted 100% recall of tumor cells which means all of the tumor cells identified by the medical professional are successfully detected, the precision of 20.6% is very low due to a large number of noise objects existing. Corresponding F_1 value is 34.2% which is also absolutely unaccepted.

Next, we perform experiments on the predictive accuracy of the three base classifiers. We use the 10 fold cross-validation testing strategy. Performance metrics of false positive

error rate, false negative error rate and average error rate, together with the scores of recall, precision and F_1 value, are computed for each fold first and then are averaged over 10 folds. The results of probabilities of false positive and false negative, probabilities of precision and recall, average error rates and F_1 values obtained for 10 splits of the database for Naïve Bayes classifier, C4.5 classifier, CBA classifier and Meta-classifier with majority vote are summarized in Tables 1, 2, 3 and 4 respectively.

The average error rate and F_1 value for Naïve Bayes classifier are 7.4% and 81.0% respectively. C4.5 has 5.7% average error rate and 86.5% F_1 value. The performance of CBA classifier leads to 6.4% of average error rate and 85.1% of F_1 value. Meta-classifier has 5.7% average error rate and 86.5% F_1 value. The C4.5 classifier has the lowest average error rate and highest F_1 value among the three base classifiers. Unfortunately, the Meta-classifier with majority vote could not provide better classification accuracy than C4.5 classier as shown in Table 5. Identification performances by using the improved image processing techniques and features mining algorithm for tumor cell identification in fluorescence cell images are summarized in Table 6. Experiment results indicate that the mining process of using extracted features rules has strong capability to distinguish tumor cells from histological noise. Satisfactory identification accuracy of 94.3% is obtained by using the improved image processing techniques and a C4.5 classifier (F_1 = 86.5%) in the present study.

*TABLE 1: Probabilities of false positive and false negative, probabilities of precision and recall, average error rates and F_1 values for the 10 splits with the **Naïve Bayes rule** classifier.*

Split *k*	**False positive rate**	**False negative rate**	**Precision**	**Recall**	**Avg. error rate**	***F*$_1$**
1	0.056	0.213	0.825	0.787	0.095	0.806
2	0.042	0.247	0.877	0.753	0.101	0.809
3	0.046	0.211	0.838	0.789	0.085	0.813
4	0.041	0.191	0.835	0.809	0.071	0.822
5	0.013	0.206	0.692	0.794	0.020	0.740
6	0.053	0.132	0.767	0.868	0.066	0.814
7	0.048	0.194	0.877	0.806	0.091	0.840
8	0.078	0.189	0.827	0.811	0.113	0.819
9	0.039	0.174	0.788	0.826	0.059	0.807
10	0.027	0.157	0.816	0.843	0.044	0.829
Avg.	0.043	0.196	0.814	0.809	**0.074**	**0.810**

*TABLE 2: Probabilities of false positive and false negative, probabilities of precision and recall, average error rates and F_1 values for the 10 splits with the **C4.5 rule** classifier.*

Split *k*	**Prob. of false positive**	**false negative**	**Prob. of precision**	**recall**	**Avg. error rate**	F_1
1	0.038	0.171	0.881	0.829	0.071	0.854
2	0.042	0.219	0.881	0.781	0.093	0.828
3	0.045	0.197	0.844	0.803	0.080	0.823
4	0.034	0.144	0.865	0.856	0.056	0.860
5	0.004	0.029	0.892	0.971	0.005	0.930
6	0.034	0.113	0.839	0.887	0.047	0.862
7	0.037	0.130	0.908	0.870	0.065	0.888
8	0.056	0.143	0.875	0.857	0.084	0.866
9	0.026	0.097	0.861	0.903	0.037	0.881
10	0.021	0.132	0.854	0.868	0.035	0.861
Avg.	0.032	0.153	0.870	0.862	**0.057**	**0.865**

*TABLE 3: Probabilities of false positive and false negative, probabilities of precision and recall, average error rates and F_1 values for the 10 splits with the **CBA rule** classifier.*

Split *k*	**Prob. of false positive**	**false negative**	**Prob. of precision**	**recall**	**Avg. error rate**	F_1
1	0.053	0.167	0.840	0.833	0.081	0.837
2	0.051	0.146	0.870	0.854	0.078	0.862
3	0.063	0.139	0.807	0.861	0.080	0.833
4	0.046	0.134	0.828	0.866	0.064	0.846
5	0.006	0.088	0.838	0.912	0.009	0.873
6	0.055	0.094	0.766	0.906	0.062	0.830
7	0.079	0.092	0.830	0.908	0.082	0.867
8	0.099	0.093	0.808	0.907	0.097	0.854
9	0.045	0.035	0.790	0.965	0.044	0.869
10	0.036	0.091	0.786	0.909	0.043	0.843
Avg.	0.051	0.114	0.816	0.892	**0.064**	**0.851**

*TABLE 4: Probabilities of false positive and false negative, probabilities of precision and recall, average error rates and F_1 values for the 10 splits with the **Meta** classifier.*

Split *k*	**Prob. of false positive**	**false negative**	**Prob. of precision**	**recall**	**Avg. error rate**	F_1
1	0.042	0.167	0.870	0.833	0.073	0.851
2	0.041	0.182	0.889	0.818	0.081	0.852
3	0.056	0.188	0.815	0.812	0.087	0.813
4	0.037	0.124	0.859	0.876	0.054	0.867
5	0.006	0.088	0.861	0.912	0.008	0.886
6	0.039	0.101	0.822	0.899	0.049	0.859
7	0.047	0.113	0.887	0.887	0.067	0.887
8	0.070	0.096	0.855	0.904	0.078	0.879
9	0.032	0.062	0.839	0.938	0.037	0.885
10	0.025	0.099	0.838	0.901	0.034	0.869
Avg.	0.038	0.130	0.853	0.878	**0.057**	**0.865**

TABLE 5: Comparison of classification performances of three base classifiers and the meta-classifiers using majority voting combination strategy.

Classification method	**System performance**	
	Average error rate (%)	**F_1 Value (%)**
Naïve Bayes classifier	7.4	81.0
C4.5 classifier	5.7	86.5
CBA classifier	6.4	85.1
Meta-classifier (Majority Voting)	5.7	86.5

TABLE 6: Comparison of identification accuracy of the improved image processing techniques (IIPT) and combination of IIPT with features mining algorithm.

Identification accuracy	**IIPT**	**IIPT + Naïve Bayes classifier**	**IIPT + C4.5 classifier**	**IIPT + CBA classifier**	**IIPT + Meta classifier**
Recall (%)	100	80.9	86.2	89.2	87.8
Precision (%)	20.6	81.4	87.0	81.6	85.3
F_1 Value (%)	34.2	81.0	86.5	85.1	86.5
Ave. error rate (%)	79.4	7.4	5.7	6.4	5.7

6. Conclusion

Advances in imaging techniques have led to large repositories of histological images in pathological studies. In this paper, we have described a real-life image mining application to the problem of tumour cell counting for fluorescence cell images. We have proposed a two-stage tumor cell identification strategy which involves segmenting objects of interest corresponding to potential tumor cells by improved image processing techniques, and identifying tumor cells using features mining algorithm. The improved image processing techniques method includes local adaptive thresholding and dynamic water immersion. Regions of interest with histological meaning are segmented from background of non-uniform illumination using a local adaptive thresholding approach. A dynamic water immersion algorithm is applied to detect individual cells in clumped cell clusters.

While the improved image processing approach is able to effectively extract reliable contours of the tumor cells, the segmentation result includes undesirable histological noise. Image processing techniques alone are unable to give accurate results for tumor cell counting. Therefore, we examine the use of features rules to distinguish tumor cells and histological noise. Meaningful features are extracted from the objects of interest and the use of extracted features rules is examined by building a number of base classifiers. A meta-classifier with majority voting strategy is also implemented. Experiment results indicate that using extracted features rules is able to distinguish tumor cells and histological noise, with an identification accuracy of 94.3%.

Acknowledgement

We would like to thank Dr Christopher Wong of Functional Genomics Laboratory, Genome Institute of Singapore, Singapore, for providing us the fluorescence cell images and valuable explanations of medical background related.

References

[1] K. L. Farina, J. B. Wyckoff, J. Rivera, H. Lee, J. E. Segall, J. S. Condeelis, J. G. Jones, "Cell motility of tumor cells visualized in living intact primary tumors using green fluorescent protein," *Cancer Research*, vol. 58, pp. 2528–2532, 1998.

[2] C.W. Wong, A. Lee, L. Shientag, J. Yu, Y. Dong, G. Kao, A.B. Al-Mehdi, E.J. Bernhard, R.J. Muschel, "Apoptosis: An Early Event in Metastatic Inefficiency," *Cancer Research*, vol. 61, pp. 333–338, 2001.

[3] J. Kittler, J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[4] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Sysst., Man Cyber.,* vol. 9, no.1, pp. 62–66, 1979.

[5] M. A. Wani, B. G. Batchelor, "Edge-region-based segmentation of range images," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 16, no. 3, pp. 314–319, 1994.

[6] R. Adams, L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 16, no. 6, pp. 641–647, 1994.

[7] J. Canny, "Computational approach to edge detection," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[8] S. Sarkar, K. Boyer, "On optimal infinite impulse response edge detection filters," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 13, no. 11, pp. 1154–1171, 1991.

[9] M.B. Jeacocke, B.C. Lovell, "A multi-resolution algorithm for cytological image segmentation," in *Proc. 2nd Australian and New Zealand Conf. on Intelligent Information Systems*, Brisbane, Australia, pp. 322–326, Nov. 1994.

[10] YM Chen, K. Biddell, AY Sun, P.A. Relue, J.D. Johnson, "An automatic cell counting method for optical images," in *Proc. IEEE BMES/EMBS*, Atlanta, Georgia, vol. 2, pp. 819, October 1999.

[11] H. S. Wu, J. Barba, J. Gil, "A parametric fitting algorithm for segmentation of cell images," *IEEE Trans. Biomedical Engineering*, vol. 45, no. 3, pp. 400–407, 1998.

[12] V. K. Kovalev, A. Y. Grigoriev, H.-S. Ahn, N.K. Myshkin, "Segmentation technique of complex image scene for an automatic blood cell counting system," in *Proc. SPIE - Medical Imaging, Image Processing*, vol. 2710, pp. 805–810, 1996.

[13] Awasthi, K. Vikas, W. Doolitle, G. Parulkar, J.G. MC Nally, "Cell tracking using a distributed algorithm for 3d image segmentation", *Bioimaging*, vol. 1, pp. 98–112, 1994.

[14] G.S. Berns, M.W. Berns, "Computer-based tracking of living cells," *Experimental Cell Research*, vol. 142, pp. 103–109, November 1982.

[15] C. G. Loukas, G. D. Wilson, B. Vojnovic, "Automated segmentation of cancer cell nuclei in complex tissue sections," in *Proc. SPIE*, Vol. 4158, pp. 188–198.

[16] D. Anoraganingrum, "Cell segmentation with median filter and mathematical morphology operation," in *Proc. Intl. Conf. Image Analysis and Processing*, Venice, Italy, pp. 1043–1046, Sept. 1999.

[17] K. Wu, D. Gauthier, M. D. Levine, "Live cell image segmentation," *IEEE Trans. Biomedical Engineering*, vol. 42, no. 1, pp. 1–12, 1995.

[18] A. K. Jain, S. P. Smith, E. Backer, "Segmentation of muscle cell pictures: A preliminary study," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol. 2, no.3, pp. 232–242, 1980.

[19] J. P. Thiran, B. Macq, "Morphological feature extraction for the classification of digital images of cancerous tissues," *IEEE Trans. Biomedical Engineering*, vol. 43, no. 10, pp. 1011–1020, 1996.

[20] K. Chan, T. W. Lee, P.A. Sample, M. H. Goldbaum, R. N. Weinreb, T. J. Sejnowski, "Comparison of machine learning and traditional classifiers in glaucoma diagnosis," *IEEE Trans. Biomedical Engineering*, vol. 49, no. 9, pp. 963–974, 2002.

[21] T. W. Nattkemper, H. J. Ritter, W. Schubert, "A neural classifier enabling high-throughput topological analysis of lymphocytes in tissue sections," *IEEE Trans. Information Technology in Biomedicine*, vol. 5, no. 2, pp. 138–149, 2001.

[22] M. L. Antonie, O. R. Za¨ýane, A. Coman, "Application of Data Mining Techniques for Medical Image classification," in *Proc. 2nd Int. Workshop Multimedia Data Mining (MDM/KDD'2001)*, San Francisco, USA, pp. 94–101, August 2001.

[23] W. Hsu, L. M. Lee, G. K. Goh, "Image Mining in IRIS: Integrated Retinal Information System," in *Proc. ACM SIGMOD*, Dallas, Texas, U.S.A., pp. 593, May 2000.

[24] L. Zheng, A. K. Chan, "An artificial intelligent algorithm for tumor detection in screening mammogram," *IEEE Trans. Medical Imaging*, vol. 20, no. 7, pp. 559–567, July 2001.

[25] L. Lam, C.Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *IEEE Trans. Systems Man Cybern*., Part A, vol. 27, no. 5, pp. 553–568, 1997.

[26] J. Kittler, M. Hatef, R. Duin, J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[27] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd Ed., Boston: Academic Press, 1990.

[28] J. R. Quinlan, *C4.5: program for machine learning*, California: Morgan Kaufmann Publishers, 1992.

[29] B. Liu, W. Hsu, Y. Ma, "Integrating Classification and Association Rule Mining," in *Proc. 4th Int. Conf. KDD*, New York, USA, pp. 80-86, 1998.

[30] P. K. Sahoo, A. K. Saltani, A. K. C. Wong, Y. C. Chen, "A survey of thresholding techniques," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 233–260, 1988.

[31] S. D. Yanowitz, A. W. Bruckstein, "A new method for image segmentation," *Computer Vision, Graphics, and Image Processing*, vol. 46, no. 1, pp. 82–95, 1989.

[32] J. Bernsen, "Dynamic thresholding of gray-level images," in *Proc. 8th Int .Conf. Pattern Recognition*, Paris, France, pp. 1251–1255, Oct. 1986.

[33] L. Vincent, P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, 1991.

[34] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[35] C.J. van Rijsbergen, *Information Retrieval*, London: Butterworths, 1979.