國立臺灣大學工學院應用力學研究所

碩士論文

Graduate Institute of Applied Mechanics
College of Engineering
National Taiwan University

Master Thesis

在非理想背景螢光影像中細胞自動計數演算法之研究

An Automatic Cell Counting Algorithm for Fluorescent Images with Non-ideal Background

李定中

Lee, Ting-Chung

指導教授：胡文聰博士、陳中明博士

Advisor: Andrew M. Wo, Ph.D.

Chung-Ming Chen, Ph.D.

中華民國 100 年 7 月

July 2011

中文摘要

細胞計量是生醫研究中不可或缺的基礎項目之一，多數的生醫研究人員必須以人工進行細胞計量與觀察(例如不同細胞數量的統計以及判斷細胞型態)，而隨著實驗需求往往耗費大量的人力與時間，在於非理想的影像背景下，不同的研究人員的主觀看法也可能影響到最後的辨認結果。

本研究針對兩種細胞株(HUVEC 和 Jurkat)之混和樣本螢光標定影像藉由影像處理技術來擷取細胞和多變量統計方法進行分析。在影像處理的部分利用 Gaussian mixture model 對影像 histogram 做分析，利用其結果取得適合的門檻值將圖片二值化，並做適當的雜訊處理後抽離成前景及背景，再對前景擷取到的物件做多張螢光訊號的特徵比對，最後以 nearest neighbor method 建立訓練資料庫，可利用特徵值對抽離的物件是否為細胞做判準，對於非理想背景下多重螢光標定細胞的影像可得到不錯且客觀的計數結果。

根據實驗結果，以 Gaussian mixture model 可有效地對影像作前景背景的分離並不受到原始影像對比度影響，訓練完的特徵對於抽取的細胞株可有 97% 準確率的辨識度，足以應付生醫研究人員細胞株計數之需求，可望未來再進一步建立病人檢體的特徵資料庫並提供醫生良好的診斷工具。

Abstract

Cell count is among the fundamental information in cytopathology and cell biology and hematology; most of the researchers count cells and observe morphology manually through microscopes to acquire statistical information. In general, cell counting is labor-intensive and time-consuming and often very operator-dependent, especially when counting cells in images with non-ideal background. And commercial products may not be able to facilitate counting for such images, either.

In this thesis, sample was prepared from mixing two kind of cell lines (HUVEC labeled with fluorescence and Jurkat not), and cells in the micrograph of samples were analyzed and recognized by image processing techniques and multivariate statistic. First, the histogram of image was classified by Gaussian mixture model method for foreground and background extraction and processed with morphological filter for noise removal. Nearest Neighbor method was used to identify targets according to their features extracted from images. The robustness of classifier was verified by k-fold cross validation. This algorithm can analyze and count cells for two fluorescence-stained cells out of non-ideal background.

Results show the image segmentation by Gaussian mixture model is virtually independent to the environmental condition of images (exposure time, contrast, and etc.) and the accuracy of recognition is around to 97% for extracting cell according to the built feature database. The algorithm serves the needs of cell counting of medical research very well.

目錄

中文摘要 I

Abstract II

目錄 III

圖目錄 V

表目錄 VII

Chapter 1. Introduction 1

1.1. Background of cell counting 1

1.2. Motivation 3

Chapter 2. Material & Method 5

2.1. Cells on a microfluidic disk 5

2.2. Procedures of cell counting from images 7

2.3. Pre-processing image of stained cells 8

2.3.1. Adjustment on image intensity 9

2.3.2. Distinguishing foreground and background from images 10

2.3.2.1. Intensity classification based on Gaussian mixture model 11

2.3.2.2. Thresholding and morphological image processing 14

2.4. Extracting objects from images 18

2.4.1. Labeling objects 18

2.4.2. Local features comparison 20

2.4.3. Extracting features 21

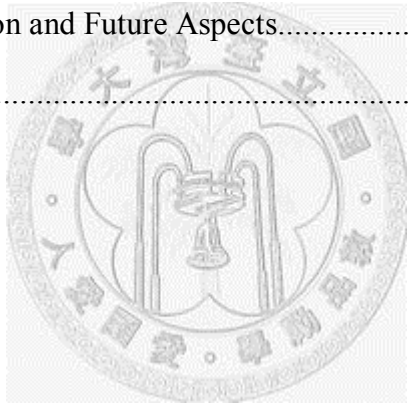
2.4.3.1. Mutual information 21

2.4.3.2. Correlation 22

2.4.3.3. Area of cell 22

2.4.3.4. Axis length of cell 23

2.5. Feature Selection and Classification 24
2.5.1. Classify Method 24
2.5.2. Nearest neighbor method 25
2.5.3. Feature selection 25
Chapter 3. Results and Discussion 26
3.1. Results of feature selection 26
3.2. Cell identification from single view field images 29
3.3. Strategies for cell recognition 31
3.4. Cell sorting for merged image 31
3.5. Region detection of inlet reservoir of the disk 33
Chapter 4. Conclusion and Future Aspects 36
Reference 37



圖目錄

Figure 2-1 Compact disk device for cell sorting (A) Layout of one sector of the disk: inlet reservoir (region A) for containing cell sample and collecting target cells for observation, connecting channel (region C, area in light green), and waste reservoir (region B). (B) Photo of the disk. 5

Figure 2-2 Compact disk, merged image and single view field diagram, every reservoir scanned with around hundred view fields, every view filed have 2 different fluorescent signal. 6

Figure 2-3 Histogram of sample image, the components of histogram concentrate in the range 50 ~ 120, the image will shows only black without intensity adjustment. 9

Figure 2-4 Adjusted image, darkest value set as the minimum intensity of the original, brightest as the maximum intensity. The cells are bright spot with shape as circle, the most area of image is background with lots noise and some impurities. 10

Figure 2-5 Gaussian Mixture model and histogram of image, the horizontal bar is intensity distribution; vertical bar is probability of intensity for each model. The sum of total area is 1, blue line is distribution of the image to fit, green lines are 3 standard Gaussian models, the red line is the Gaussian mixture model. The figure indicate 2 models fitting for background (darker intensity) and 1model fitting for foreground (brighter intensity). 12

Figure 2-6 Result of cluster of image, this figure the most of pixel is background (red color), and the blue color is foreground. 13

Figure 2-7 Threshold at third cluster, most of foreground elements are segmented with a small amount of noise, and some targets were segmented into many fragments. To remove noise and connect fragments of targets, we applied close operator first then open operator [5]. For convenient, the image was cut from origin one to display. 15

Figure 2-8 Operator of 3x3 window 16

Figure 2-9 Process of fragments merging and noise removal (a) example window of image (b) image after applied close operator, both 2 target were processed into integrated targets(c) image after applied open operator, the pixel area within 2x2 were gone. 17

Figure 2-10 Target mask with label, the targets was applied area filter limit in 10 to 1000 pixels. 18

Figure 2-11 Original Image (red fluorescence) with label. 19

Figure 2-12 Targets within local window of label 1, 2, 4, the left one of window is cutting from red fluorescent image, right one is cutting from green fluorescent image. 20

Figure 2-13 Calculating area in different method(a) Threshold result of cell stained with weak fluorescent signal (b) Convex area of cell..... 22

Figure 2-14 Shape of a cell..... 23

Figure 2-15 Example of k-Nearest Neighbor 25

Figure 3-1 Example of two dimensional feature with classified labeled plot. 26

Figure 3-2 Classified boundary of training data by nearest neighbor method 30

Figure 3-3 A rectangle region image combine with 3 continuous view fields 32

Figure 3-4 The same cells appear in different pictures due to the overlap of scan system..... 32

Figure 3-5 Threshold images of each fields, these images indicate the process acquire similar segmental result of the same cells..... 32

Figure 3-6 Result of union operation, the image mask merged according the position from scanner or computation, the repeated target were united. 33

表目錄

Table 3-1Accuracy of classification with single feature 27

Table 3-2 Accuracy of classification with various feature combination 28

Table 3-3 Example of correlation and area chart 29

Chapter 1. Introduction

1.1. Background of cell counting

Medical image processing has become important in clinic with the progresses of image theory and technology in recent years, and the application of medical image is very wide. In a broad sense, an image, which contains physical information of human and is helpful for medical works or researchers, is a medical image, such as X-ray film image, magnetic resonance imaging (MRI) and computed tomography (CT), which are all known to medical image. Some researchers has developed computer-aided diagnosis (CAD) and computer-assisted surgery(CAS) technology by combining computers and medical images in recent years, and image segmentation is one of the most important procedure for all of the technologies

There has been an increasing need to quantify medical sample for biomedical research; there are various ways to quantify cell numbers with fluorescent markers by fluorescence microscopy. Compared to quantitative polymerase chain reaction (Q-PCR) [1], the method based on microscopy offers very good advantages of visual feedback, the biases of quantification can be minimized by cautious examination of the samples, Q-PCR quantifies the numbers of the target genes in samples. In many cases, gene numbers are not convertible to cell numbers because the target gene number per microbial genome is unknown, and the number of genome copies per microbial cell can vary in different growth phases [2].

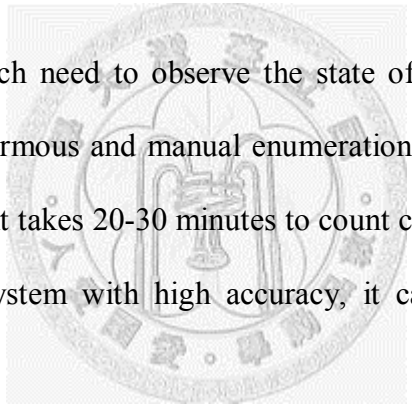
In the overview of selected fluorescence labeling quantification methods, the most used method is fluorescence in situ hybridization (FISH)[2],[3], However, one disadvantage of FISH is that it is not suitable to differentiate microbes below the species level[3]. Visual counting of labeled cells is a simple quantification method with manpower and time consuming especially in complex environmental samples. Generally, the samples are applied on a microscope slide and the cells were labeled, but in some cases, the samples were applied on low-cost device, the result is operator dependent. In order to get an objective cells count, many systems or algorithms were developed to solve the quantification problems by images analysis.

Image segmentation for cells is a procedure which extracts some components that we are interested in from an image. Because of the needs of analyses such as description, recognition and others, all of those analyses need the divided image from image segmentation, therefore, image segmentation is the most important preprocess in image processing.

1.2. Motivation

There are many developed methods of segmentation recently, such as edge detection, thresholding, region growing and active contour. Generally, the complicated algorithm provides better result of segmentation, there are cases of applications for medical images by various segmentation algorithms, but no algorithm does perfect image segmentation due to the differences between medical images and general images, such as the high noise ratio in medical images, lower contrast and complexity of geometrical shape. For these reasons above, medical research in image segmentation faces substantial challenge.

Most clinical research need to observe the state of cell from patients, cells count in the blood is enormous and manual enumeration is labor intensive, a 2ml blood sample from patient takes 20-30 minutes to count cells, if we can develop an automatic recognizing system with high accuracy, it can reduce the burden of clinic researchers.



However, there are some problems in cell image segmentation; it's difficult to do automatic cell-counting due to the shape variation, intensity variation of fluorescent cell, complexity of the background of the image. Due to the different condition or observer the result of medical image diagnosis becomes different, if there is a standard for analysis or assist would be helpful.

In this thesis, we tried on Thresholding method with computation of Gaussian probability model of images to get good result of segmentation. And acquire information of segmented objects to do recognition by Nearest Neighbor Method,

because the accuracy of recognition result could be enhanced with training database.

Chapter 2. Material & Method

2.1. Cells on a microfluidic disk

The test condition base on a compact disk device as shown in Figure 2-1 , the device provides several chamber to observe, the observe field in the device is a ellipse-like shape, and the view fields up to 100 views under microscope.

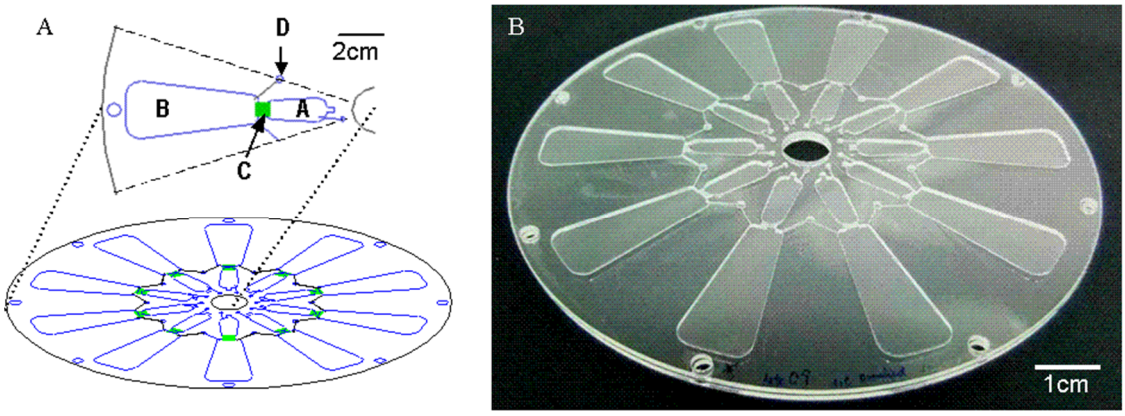


Figure 2-1 Compact disk device for cell sorting (A) Layout of one sector of the disk: inlet reservoir (region A) for containing cell sample and collecting target cells for observation, connecting channel (region C, area in light green), and waste reservoir (region B). (B) Photo of the disk.

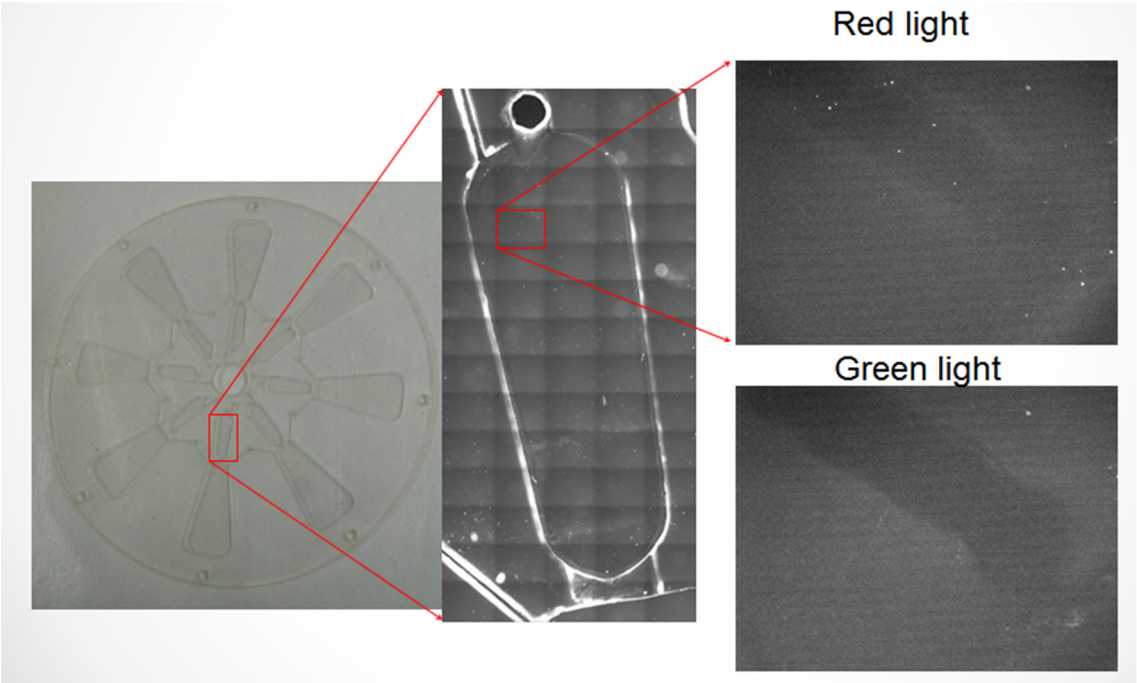


Figure 2-2 Compact disk, merged image and single view field diagram, every reservoir scanned with around hundred view fields, every view filed have 2 different fluorescent signal.

2.2. Procedures of cell counting from images

The data were collected using experimental disk or chip filled with fluorescent cell line, there are two type of dye can help us distinguish cell and impurities under the microscope. Store the images of each parts of the view field of the disk after manual check for focal distance and exposure time.

Matlab was used to pre-process images of stained cells. First, the interested information of the image was emphasized by adjusting on image contrast, then calculating the fitting Gaussian mixture model for the histogram to distinguish foreground from background. Next, the images were binarized according to the result of foreground-background clusters, and the binarized images were processed through morphological filters to reconstruct shape of cells and remove noise.

In the part of extracting objects from images, the processed image was labeled with numbers and the objects position was recorded. Each object was cropped accordingly, then getting the local features, such as mutual information, correlation, p-value, or morphological information of cells.

Finally, nearest neighbor method was used to classify objects. The model was trained and individual cell images were tested, then computing the accuracy of recognition.

2.3. Pre-processing image of stained cells

To get better result, the image needs to adjust via pre-processing before image segmentation or characteristic extraction. Generally, the complicated algorithm provides better result of segmentation, but complicated algorithm including more parameters, most of time, the adjustment of parameters is not regular and time-consuming, the user need to try and error for adjustment on algorithm, it's not suit for practical application.

A digital image could be defined as a numeric matrix; the value $f(x,y)$ at point (x,y) is the intensity of image, the brighter point with higher intensity, gray scale is the intensity of mono color image, and general color image usually combine with 3 mono color images (red, blue green), the function of digital image could express as following expression.

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & f(0,2) & \cdots & f(0,N-1) \\ f(1,0) & f(1,1) & f(1,2) & \cdots & f(1,N-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(M-1,0) & f(M-1,1) & f(M-1,2) & \cdots & f(M-1,N-1) \end{bmatrix}$$

This expression indicates the intensities and position of every point of a digital image.

2.3.1. Adjustment on image intensity

The histogram of digital images is a discrete function indicates the numbers of pixels to intensity; we can emphasize or analyze certain information of images by adjusting the histogram.

The sample image and its histogram shown as Figure 2-3 and Figure 2-4, this image is 16 bit gray scale (range of intensity is 0~65535) format, but the effectual information only 8-12 bit (range of intensity is 0~255 or 0~4096) depends on the picture exposure time or other condition in the most of time, so the first step is to extract useful information and exclude garbage data from raw image. To get the effectual information, we extract the available part of histogram of the image. This step not only make the image more clear for display but reduce the size of data, less data size reduces the compute times.

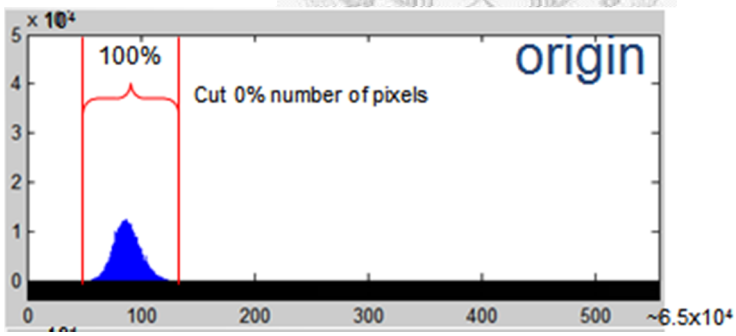


Figure 2-3 Histogram of sample image, the components of histogram concentrate in the range 50 ~ 120, the image will shows only black without intensity adjustment.

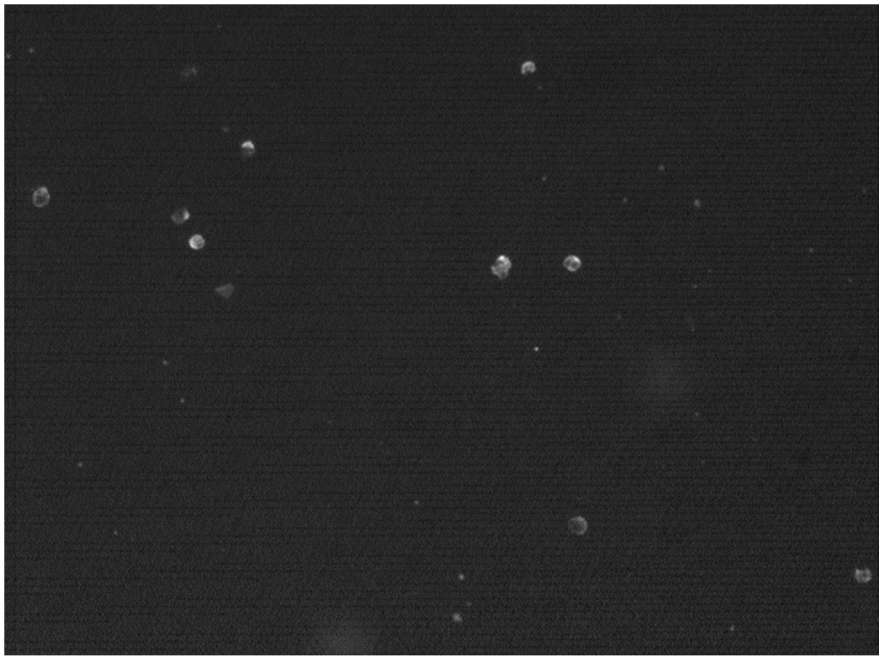


Figure 2-4 Adjusted image, darkest value set as the minimum intensity of the original, brightest as the maximum intensity. The cells are bright spot with shape as circle, the most area of image is background with lots noise and some impurities.

2.3.2. Distinguishing foreground and background from images

Before segmentation, in order to extract objects from background ,the images need to be processed for characteristic enhancement, most of cases of image segmentation were using thresholding, the result of segmentation depends on the value choose for threshold.

We suppose the distribution components of image are Gaussian distribution, to divide the foreground and background, we simulate several various Gaussian models to fit the histogram of images. To segment the image, we do threshold

process the after the mean and standard deviation of Gaussian model were computed.

After classified the intensity for foreground and background, threshold the images into binary images, then process with morphological filters to get good images to segment.

2.3.2.1. Intensity classification based on Gaussian mixture model

To classify the intensities of images to foreground and background, the Gaussian mixture model(GMM)[4][5][6][7] method was used in fitting intensity distribution, the Gaussian probability density function as equation (2.1) , Gaussian mixture model function as equation (2.2), the sum of weighting factor is 1 as equation (2.3).

The sample image shown as Figure 2-4 ,its histogram is the blue line shown in Figure 2-5,and the red line is the Gaussian mixture model fitting to image histogram, the green lines are 3 individual Gaussian models.

$$g(x;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \quad (2.1)$$

$$p(x)=\alpha_1 g(x;\mu_1,\Sigma_1)+\alpha_2 g(x;\mu_2,\Sigma_2)+\alpha_3 g(x;\mu_3,\Sigma_3) \quad (2.2)$$

$$\alpha_1+\alpha_2+\alpha_2=1 \quad (2.3)$$

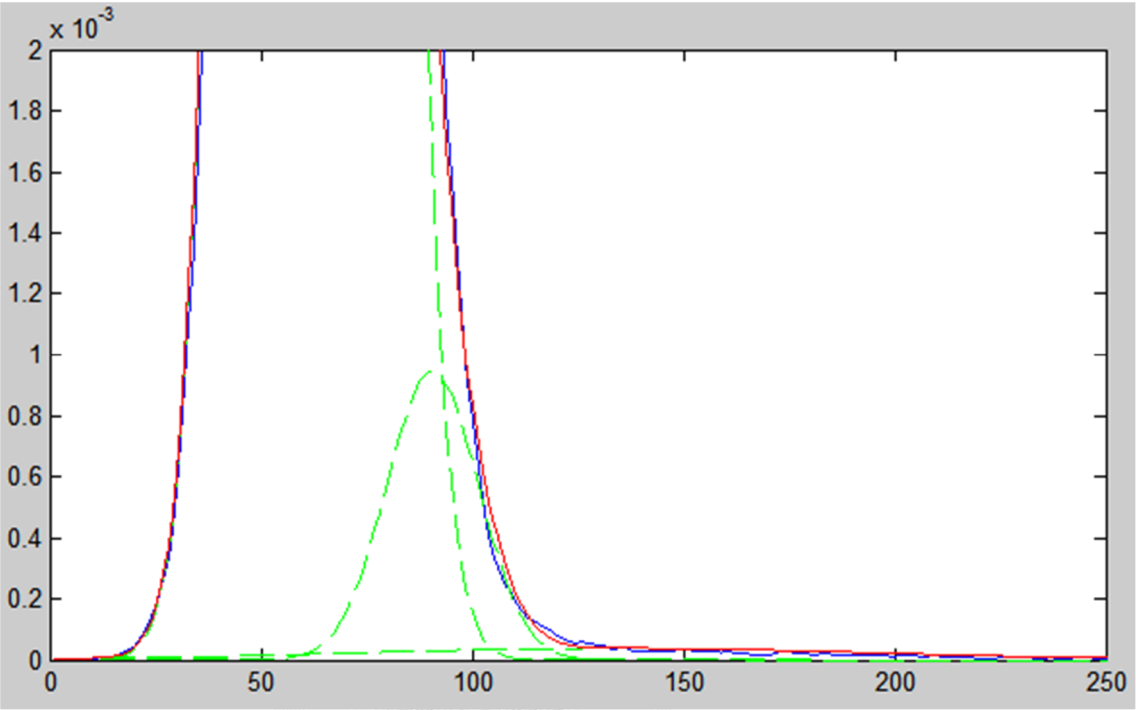


Figure 2-5 Gaussian Mixture model and histogram of image, the horizontal bar is intensity distribution; vertical bar is probability of intensity for each model. The sum of total area is 1, blue line is distribution of the image to fit, green lines are 3 standard Gaussian models, the red line is the Gaussian mixture model. The figure indicates 2 models fitting for background (darker intensity) and 1model fitting for foreground (brighter intensity).

After the computation of the distribution of Gaussian mixture model, the maximum probabilities of intensity could be found, then cluster and label the possible cluster for every pixel. Then we gain the visualized distribution on image shown as *Figure 2-6*

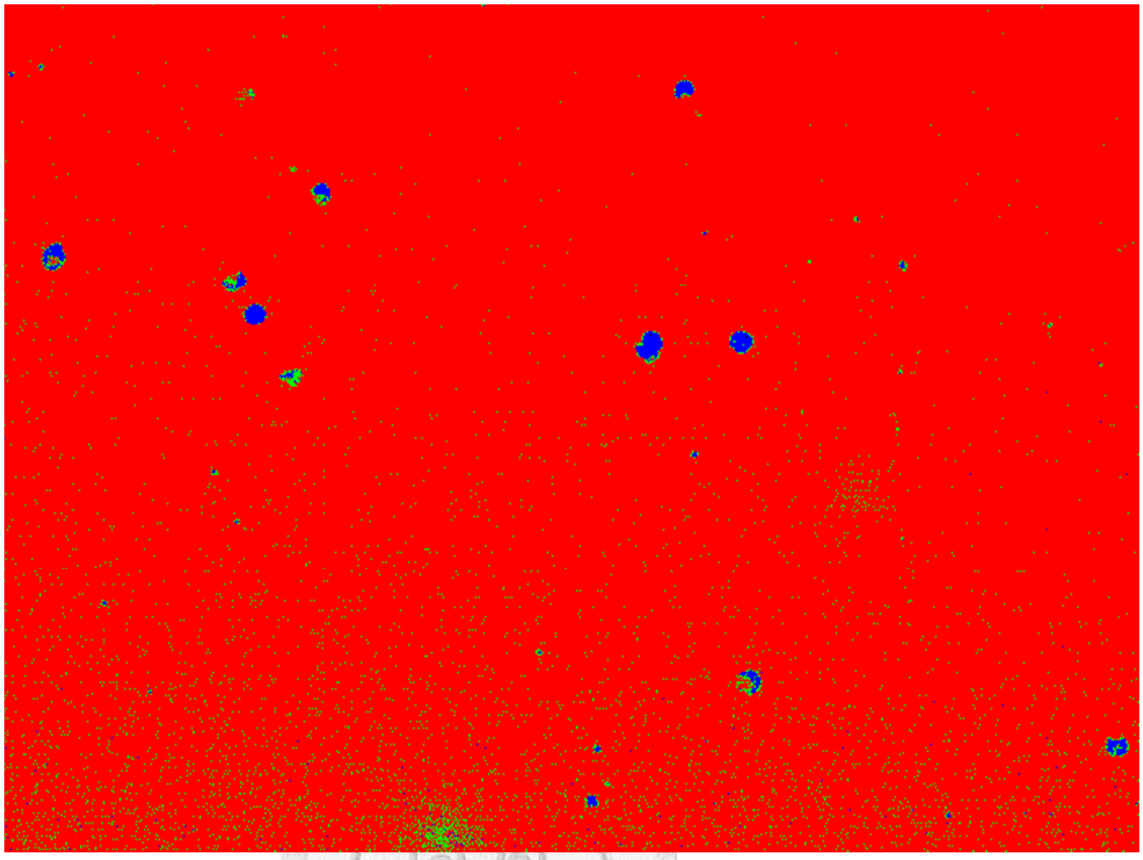


Figure 2-6 Result of cluster of image, this figure shows the most of pixel is background (red color), and the blue color is foreground.

The colorized image for cluster shown as Figure 2-6, this image shows 3 clusters perfectly, the red for background occupied most area, the blue for fluorescence targets and green for microscopy noise or impurities. In the method of Gaussian mixture model approach, the adjusting on parameter (for example, the exposure time of images) is virtually independent to contrast of images.

2.3.2.2.Thresholding and morphological image processing

After cluster foreground and background of the image, we need to decide a threshold value to make the image binary for dividing foreground (fluorescent cells, noise) and background (device background, others),where the threshold value must be a fitting value for segmentation. The quality of image segmentation rest on the threshold image, and the result, in order to exclude some noise and connect some fragments of targets, the image will be applied close and open operator.

Thresholding

$$g(x,y) = \begin{cases} 1 & if\ f(x,y) > T \\ 0 & if\ f(x,y) \leq T \end{cases}$$

First, we threshold image at third cluster (blue color in Figure 2-6), the processed images shown in Figure 2-7, from the picture, we can found some noises around 1 pixel spread out the background. Some of the targets were segmented to fragments.

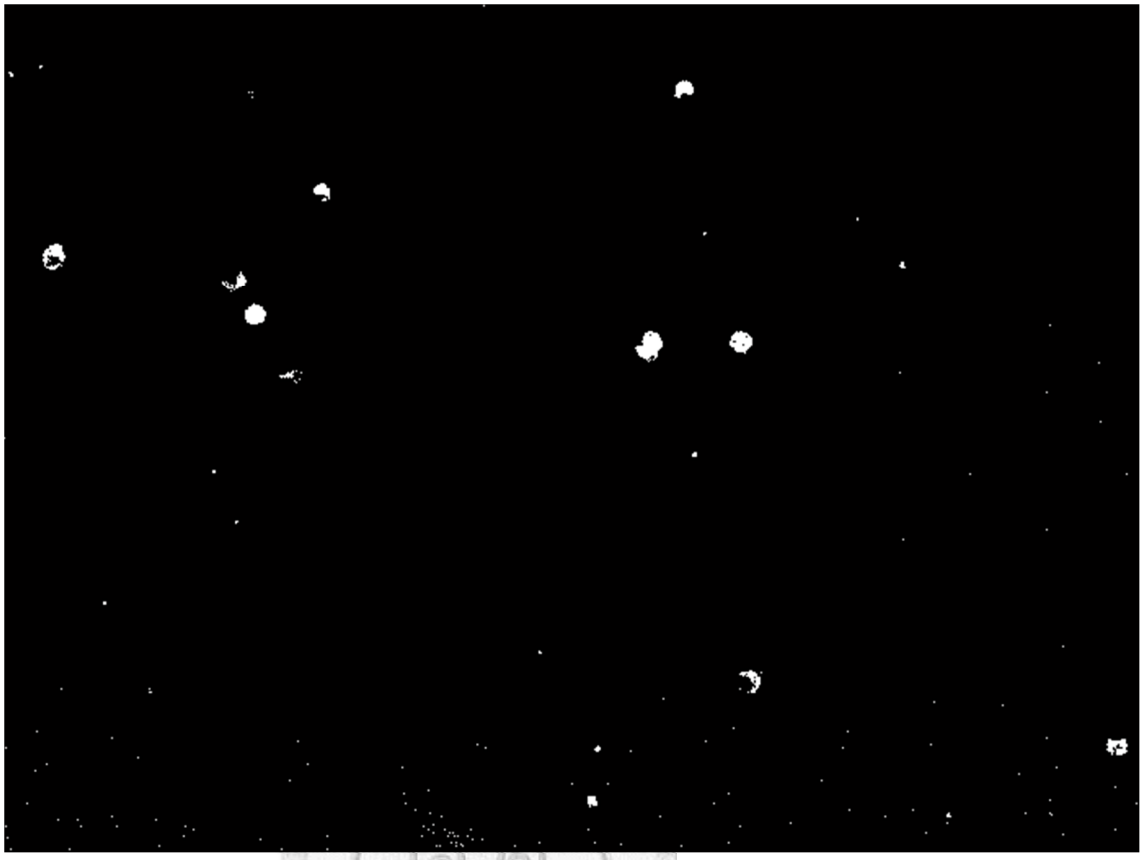


Figure 2-7 Threshold at third cluster, most of foreground elements are segmented with a small amount of noise, and some targets were segmented into many fragments. To remove noise and connect fragments of targets, we applied close operator first then open operator [8][9]. For convenient, the image was cut from origin one to display.

The open and close filters were applied with a 3x3 window operator, open filter can erode the image small than the applied filter, close window connect object with small gap.

0	1	0
1	1	1
0	1	0

Figure 2-8 Operator window of 3x3

Dilution

$$A \oplus B = \left\{ z | \left(\hat{B} \right)_z \cap A \neq \phi \right\}$$

Erosion

$$A \ominus B = \left\{ z | \left(\hat{B} \right)_z \subseteq A \neq \phi \right\}$$

Open operation

$$A \circ B = (A \ominus B) \oplus B$$

Erosion first then dilution

Close operation

$$A \cdot B = (A \oplus B) \ominus B$$

Dilution first then erosion

(a)

(b)

(c)

Figure 2-9 Process of fragments merging and noise removal. (a) Example of window of image. (b) Image after applied close operator. Both targets were processed into integrated targets. (c) Image after applied open operator, the objects with area within 2x2 were gone.

2.4. Extracting objects from images

In this part, to get information for the following analysis the image was divided into many local images with objects individually, the threshold value was estimate by Gaussian mixture model.

2.4.1. Labeling objects

The threshold image was labeled numbers by connected-component based on 4-connectivity [8], this step record the object information of positions, area, axes lengths for following manual check.

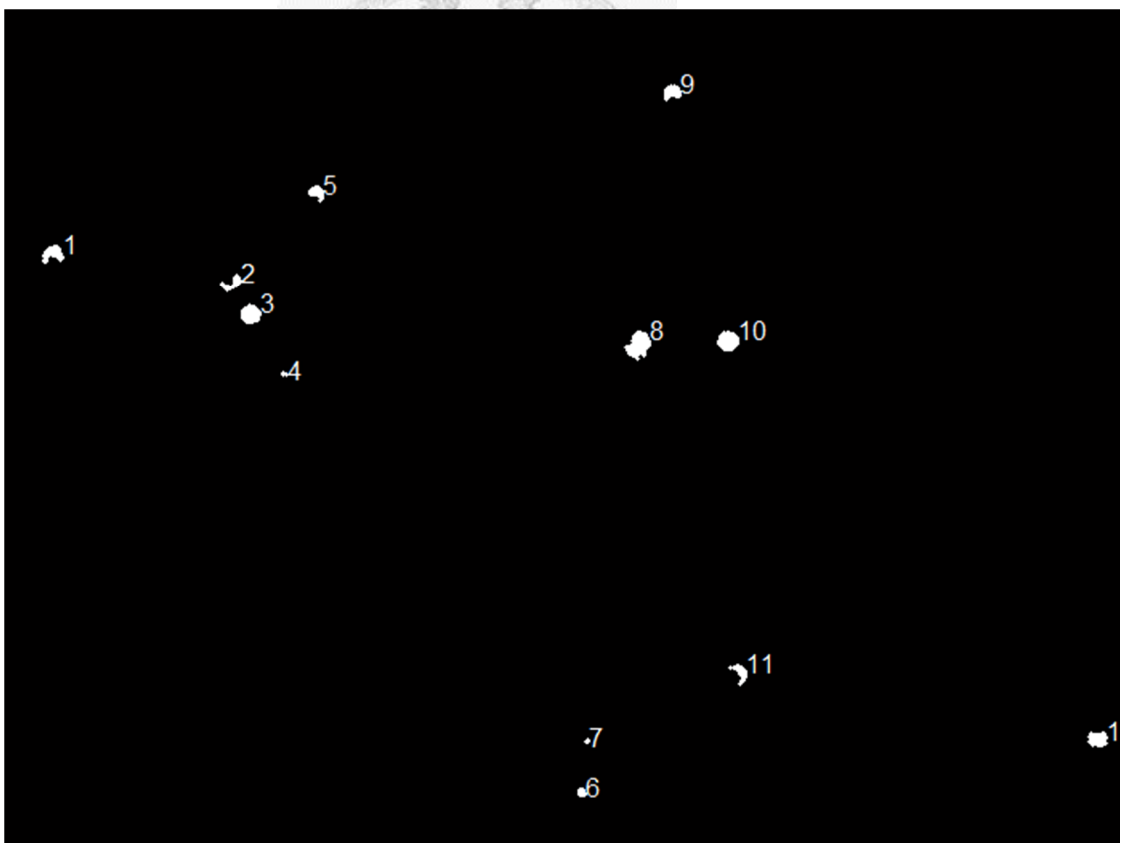


Figure 2-10 Target mask with label, the targets was applied area filter limit in 10 to 1000 pixels.

The gray scale image also applied with the label for manual checking, these data will be used in classification section.

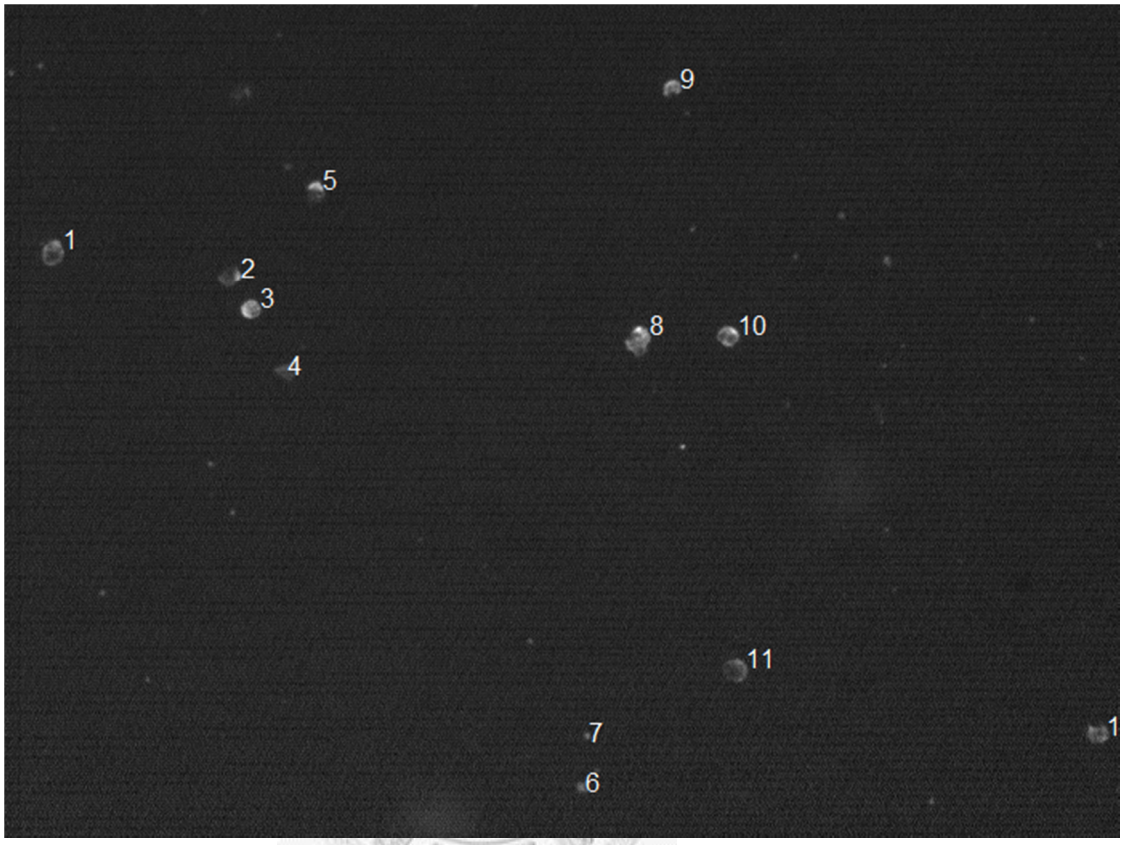


Figure 2-11 Original Image (red fluorescence) with label.

2.4.2. Local features comparison

In this step, we calculate the areas of every labeled target, and then select twice size of area of target as a window, compare information of the windows from original images with different fluorescence.

The compare window shown in Figure 2-12, it's clear that #1 and #2 only have red fluorescent signal, but #4 have both red and green, so we can sure that #4 target is impurity or noise.

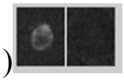
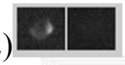
(#1)  (#2)  (#4) 

Figure 2-12 Targets within local window of label 1, 2, 4, the left one of window is cutting from red fluorescent image, right one is cutting from green fluorescent image.

2.4.3. Extracting features

The features of image is the key to determine a system can be used in clinical research, the features including morphological features and texture features, for the application on cell counting, the quality of stained cells are not regular, and the condition of exposure time is not stable.

In this thesis, we want to find some features which less influence with environmental condition to raise the accuracy of recognition.

2.4.3.1.Mutual information

The relativity of intensity between images was calculated by mutual information method [10], the similarity can be identified according to mutual information

Mutual information $I(x,y)$ of vector x and y,

$$I(x,y) = I_{xy} = H_x + H_x - H_{xy}$$

$$= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{xy}(x,y) log \frac{f_{xy}(x,y)}{f_x(x)f_y(y)} dx\, dy,$$

where $H(x) = H_x = -\int_{-\infty}^{\infty} f_x(x) log f_x(x) dx$ is entropy of image or vector x, and

$$H(x,y) = H_{xy} = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{xy}(x,y) log f_{xy}(x,y) dx\, dy$$

2.4.3.2.Correlation

Calculate the correlation between two objects; if the correlation closes to zero means two object are different, contrarily, they are similar.

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}}$$

2.4.3.3.Area of cell

(a)Filled area of target (b)Convex area

Figure 2-13 Calculating area in different method. (a) Threshold result of cell stained with weak fluorescent signal, (b) convex area of cell

Figure 2-13 shows that that area of an extracted object sometimes does not express the shapes of cells correctly when the cells were weakly stained. In order to solve this problem, we use convex area to replace filled area. In this way, the area will close the normal cell size, and noise size will not change, the cells can be distinguished from noise.

2.4.3.4.Axis length of cell

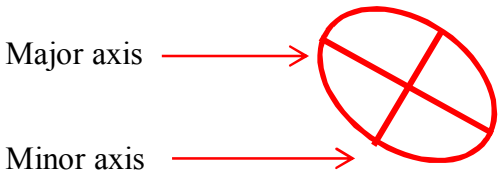


Figure 2-14 Shape of a cell

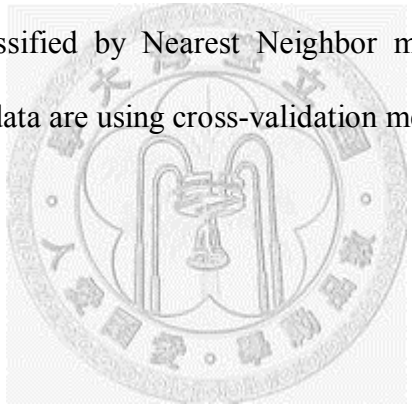
Figure 2-14 shows the axes length of a cell. Normally, the shape of a cell is ellipse or circle, the length of major axis (the maximum length of shape) and minor axis (axis perpendicular to major axis) are close, but noise or impurities are not, the ratio of major axis and minor axis is a reference to recognize cells.

2.5. Feature Selection and Classification

There are many features data after extracting features, to classify cells and impurities by the most influential features, if the data was classified with all features, the result will be interfered, it's necessary to select feature for classification.

2.5.1. Classification Method

There are various ways for the feature classification, expert system, decision tree [11], support vector machine (SVM) [12], neural network and others. In this thesis, the data was classified by Nearest Neighbor method [13], besides; the training data and testing data are using cross-validation method.



2.5.2. Nearest neighbor method

To use Nearest Neighbor method to classify object, the data need to normalized in a range [14][15], the training sample features are vectors in a multidimensional space, the algorithm find the nearest distance between training sample and the testing sample to classify samples.

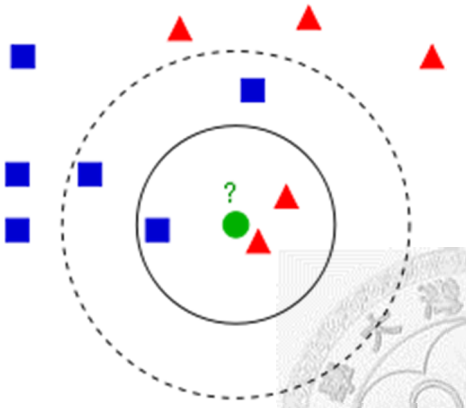


Figure 2-15 Example of k-Nearest Neighbor

2.5.3. Feature selection

There are two way to select features, forward feature selection and exhausted feature selection. The exhausted selection tests all the combination of features and get best combination for classification but spend more time, and forward feature selection is more economical way to find useful features.

Chapter 3. Results and Discussion

3.1. Results of feature selection

To find best combination of features for classification and verify the reliability, this experiment tests the performance by k-fold cross-validation with forward feature selection.

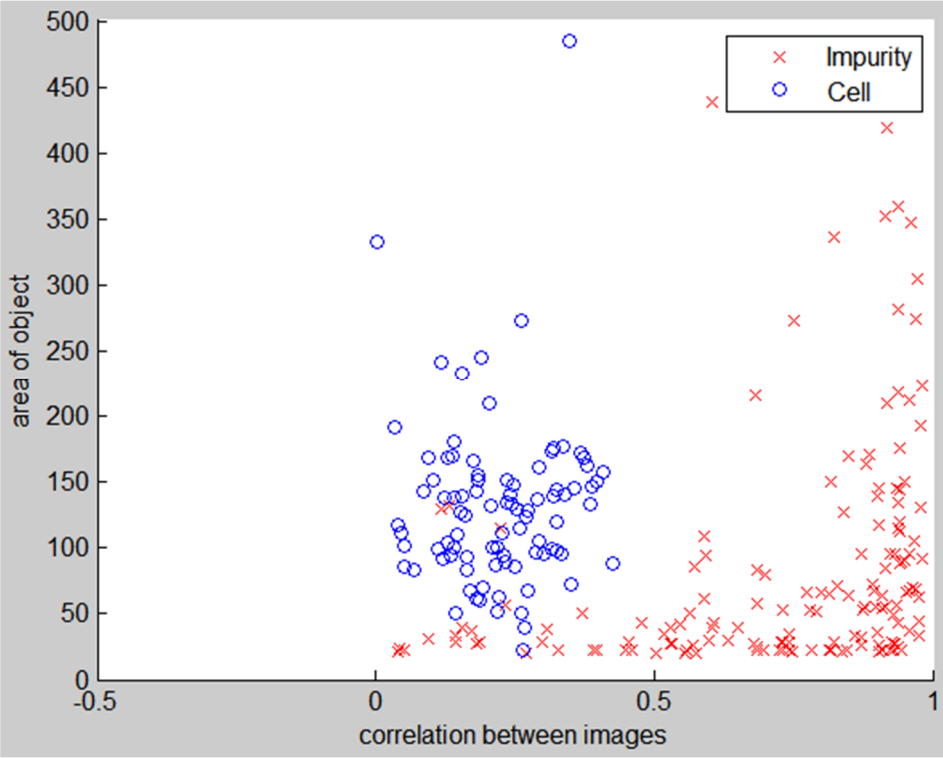


Figure 3-1 Example of two dimensional feature with classified labeled plot.

	True Positive	False Positive	True Negative	False Negative	All	recall	Precision	Err of cell number
Correlation	5.6%	9.1%	73.4%	11.9%	79.0%	38.4%	32.2%	-16.3%
Area	2.3%	6.1%	76.3%	15.2%	78.7%	27.5%	13.2%	-51.9%
convex area	4.1%	5.8%	76.6%	13.5%	80.7%	41.0%	23.1%	-43.6%
axial ratio	9.7%	10.9%	71.6%	7.8%	81.3%	47.3%	55.4%	17.2%
mutual information	10.8%	12.9%	69.6%	6.7%	80.4%	45.6%	61.6%	35.1%

Table 3-1 Accuracy of classification with single feature

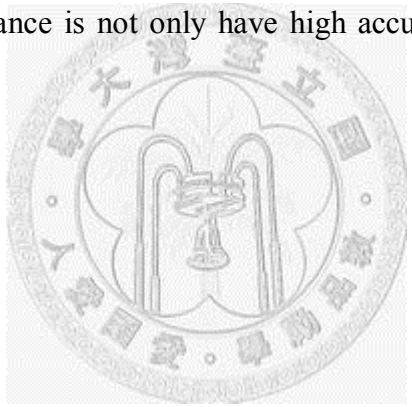
	Meaning.
True positive	The accuracy of cell recognized as cell
False positive	The accuracy of impurity recognized as cell
True negative	The accuracy of impurity recognized as impurity
False negative	The accuracy of cell recognized as impurity
All	The accuracy of total recognition
Recall	The fraction of the targets that are relevant to the query that are successfully retrieved
Precision	The fraction of retrieve targets that are relevant to the search
Error of cell number	The error of the number between positive and cell

From Table 3-1, the best performance feature for classification is axial ratio of object, the result shows that axial ratio filter performs best accuracy and recall and second rank precision. Although axial ratio feature has the best performance in single feature comparison, the correlation obtain less error of cell numbers, so correlation feature was choose for the base feature for combination in next feature selection.

	True Positive	False Positive	True Negtive	False Negative	All	recall	Precision	Err of cell number
Correlation area	16.3%	1.3%	80.1%	2.4%	96.4%	92.5%	86.7%	-5.4%
Correlation convex area	15.8%	1.8%	79.8%	2.7%	95.5%	89.3%	84.7%	-4.4%
Correlation axial ratio	9.8%	7.7%	77.1%	5.4%	86.9%	61.0%	69.1%	18.1%
Correlation mutual information	8.9%	8.7%	75.6%	6.8%	84.5%	55.1%	61.1%	14.8%

Table 3-2 Accuracy of classification with various feature combination

From Table 3-2, the best feature combination for classification is correlation and convex area (according to the error range of cell number), the table shows that all performances of combinations are better than single feature. The best feature combination for performance is not only have high accuracy, but high recall and precision around 90%.



3.2. Cell identification from single view field images

The output result shown as Table 3-3, with the parameters (area of cell and correlation between different fluorescence), we can use nearest neighbor method to get class probability distribution.

Correlation	Area	Group
0.129511	166	Cell
0.407957	157	Cell
0.934183	149	Impurity
0.984538	91	Impurity
0.262623	81	Cell
0.929376	23	Impurity
0.741446	21	Impurity
0.839475	19	Impurity
0.20483	198	Cell
0.696385	42	Impurity

Table 3-3 Example of correlation and area chart

To sorting the data, we use linear, quadratic and nearest neighbor classify method to classify the characteristics, the results shown as Figure 3-2, the visualized result of classification can be used to examine the distribution of targets in the feature space.

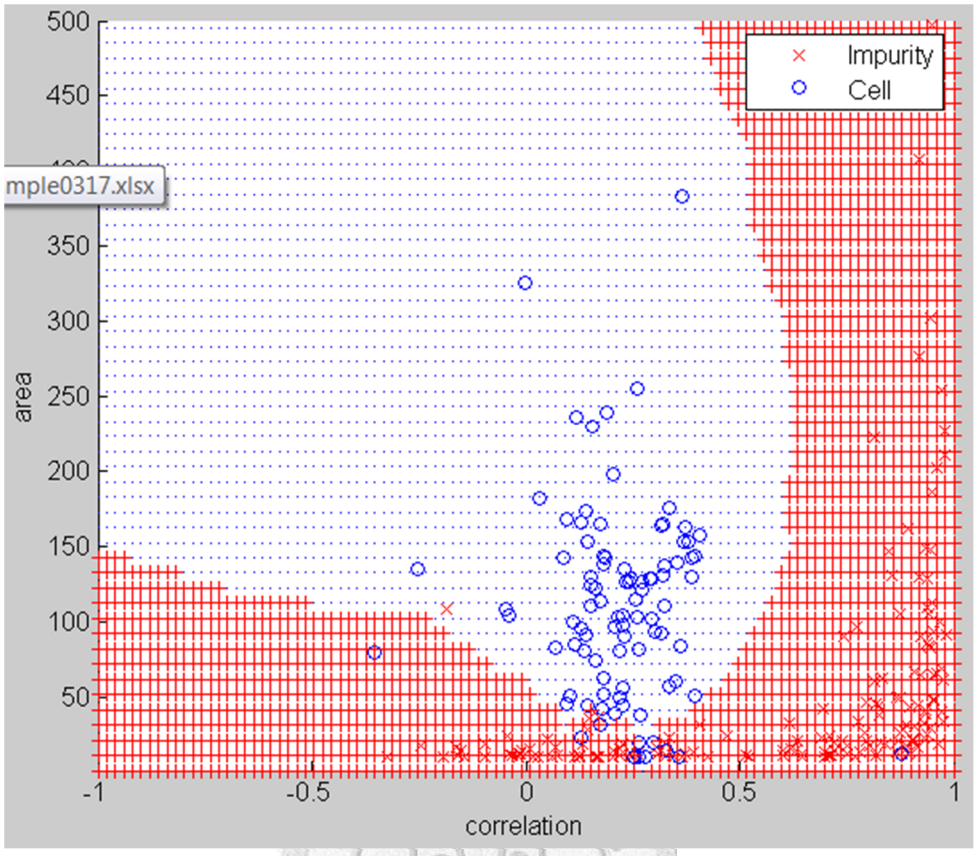


Figure 3-2 Classified boundary of training data by nearest neighbor method

3.3. Strategies for cell recognition

In General, the method used for recognizing multi-fluorescence stained cell is thresholding image by average background or adjust manually.

This system determine background by distribution of image intensity, the threshold value was adjusted automatically, the threshold value determination is the important part in the most of time, a bad threshold process for images will result in the failure on objects and features extraction.

In 2.3.2.1, we can see the distribution of clustering models, in this way, the threshold efficient is better than traditional method and avoiding try and error.

3.4. Cell sorting for merged image

In practical situation of cell counting, it's usually that observed region of slides or devices with large area. For instance, the size of counting region shown as Figure 3-3 is a rectangle view fields, and image scanned into 3 images with different position. To merge the images correctly, each image scanned with overlap margin to check and fitting adjacent images. The process of merged image should acquire the relative position between images before the combination.

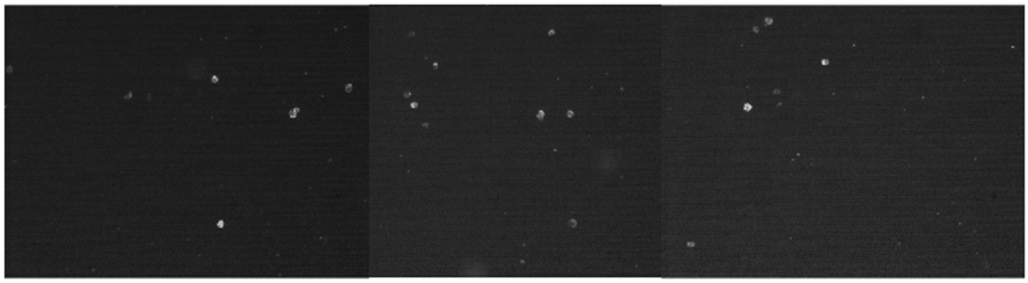


Figure 3-3 A rectangle region image combine with 3 continuous view fields

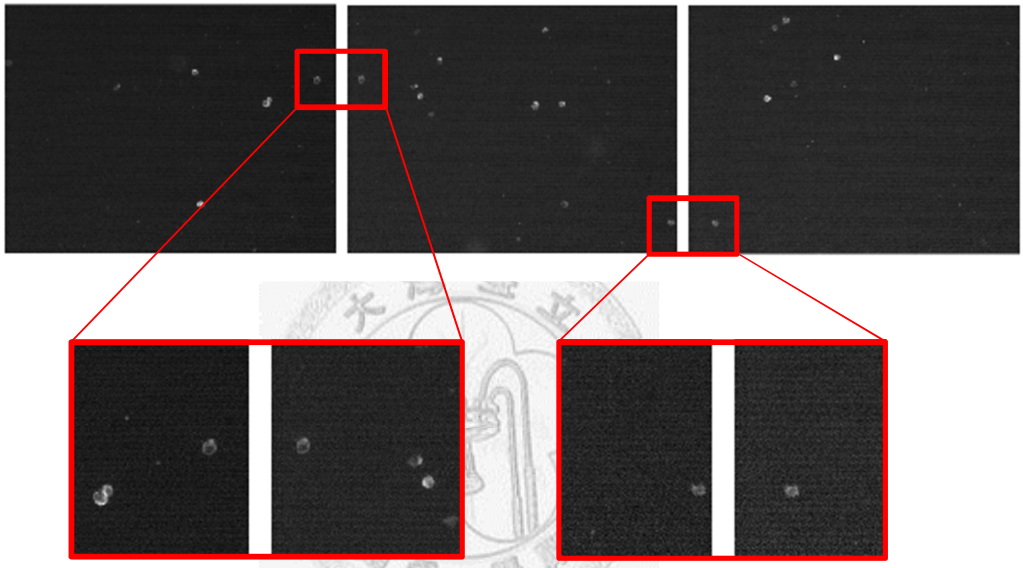


Figure 3-4 The same cells appear in different pictures due to the overlap of scan system.

Figure 3-5 Threshold images of each fields, these images indicate the process acquire similar segmental result of the same cells.

Figure 3-6 Result of union operation, the image mask merged according the position from scanner or computation, the repeated target were united.

3.5. Region detection of inlet reservoir of the disk

The ROI (Region of interest) of the shape of samples is not regular sometimes; the computation for the area of images outside ROI is times consuming with higher fail results.

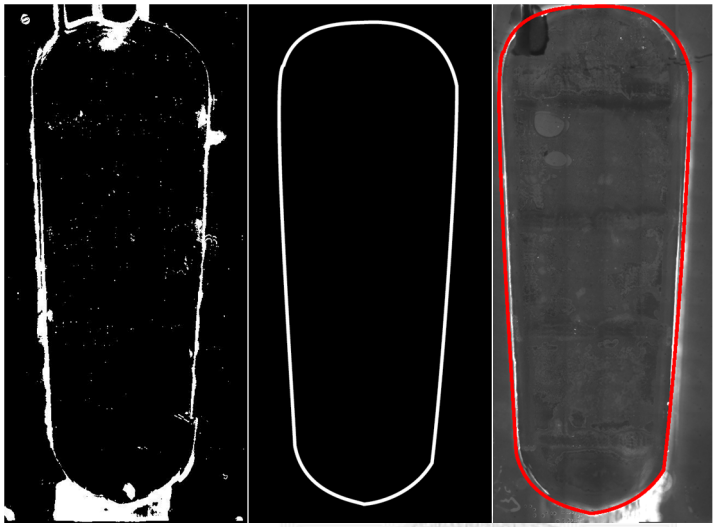
To remove the redundant part of image, I prepare a vector model to compare with threshold images. Compute and find out the position which most overlapping and fitting.

$$\text{Max}(\boldsymbol{\Sigma}[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \cap \boldsymbol{m}(\boldsymbol{x},\boldsymbol{y})])$$

$\boldsymbol{f}$:original image $\boldsymbol{m}$: model image

The purpose of this step is reducing the time of computation by avoiding process the part of image without cells. This step could save 30% time for average. Generally, boundary detection is an extend process for merged image, the performance time is longer than single view field images. The images can be

shrinking to a small size for faster performance time, the error increasing with the resize ratio.



Take an overview of 3.4 and 3.5 the procedure can be revise to following step:

1. Region detection for device boundary.
2. Adjusting histogram of image of single field.
3. Gaussian mixture model for dividing foreground and background for each single view field.
4. Merged image combination.
5. Applying morphological filters
6. Feature extraction
7. Classification

In the step 2 and 3, the single view field images, rather than entire view images, will be processed, for the local threshold (threshold to single view field) has better performance than thresholding to merged image, because each single view field images has its background intensity distribution and it is different from entire background intensity distribution. After the single view field images processed as binary mask images, the binary images were combined as merged image in step 4 for the subsequent procedures.

Chapter 4. Conclusion and Future Aspects

In this thesis, a system was developed to count cell with image processing technic and expert system training to get fine results. In the part of image processing, the samples analyzed and distinguished foreground and background without optimizing complicated parameters manually.

The accuracy of results is around 95% for extracting cell in the 10-fold cross validation test, and the recall and precision were around to 90%, it is sufficient for the needs of cell counting of medical research, the calculating method of cell area and the numbers comparison of nearest neighbor method is two of the most influential part, the calculating way of area size is helping for the dispersing cells and impurities data. In this experiment, I only use 2 features (area of cells and correlation between local images), the results show that the performance of feature combination is better than single feature ,this system can be enhance with more influential feature of cells.

The future work will involve more feature extraction of cells, and combination of cell recognition and merged images region detection, for the furthermore, recognize the type of cells by morphological features. The complete system can implement by integrating the processes of region detection and merged image in the future work.

Reference

[1] L. Bjerrum, T. Kjaer, and N.B. Ramsing, "Enumerating ammonia-oxidizing bacteria in environmental samples using competitive PCR," *Journal of microbiological methods*, vol. 51, 2002, pp. 227–239.

[2] M. Wagner, M. Horn, and H. Daims, "Fluorescence in situ hybridisation for the identification and characterisation of prokaryotes," *Current Opinion in Microbiology*, vol. 6, Jun. 2003, pp. 302-309.

[3] H. Daims and M. Wagner, "Quantification of uncultured microorganisms by fluorescence microscopy and digital image analysis," *Applied Microbiology and Biotechnology*, vol. 75, Mar. 2007, pp. 237-248.

[4] J.A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, 1998, p. 126.

[5] A.K.C. Wong and P.K. Sahoo, "A gray-level threshold selection method based on maximum entropyprinciple," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, Aug. 1989, pp. 866-871.

[6] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using Expectation-Maximization and its application to image querying," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 24, 1999, p. 1026--1038.

[7] H. Hong and D. Schonfeld, "Maximum-entropy expectation-maximization algorithm for image reconstruction and sensor field estimation," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 17, Jun. 2008, pp. 897-907.

[8] R. Gonzalez, *Digital image processing*, Upper Saddle River N.J.: Pearson/Prentice Hall, 2010.

[9] P. Dodwell, *Visual pattern recognition*, New York: Holt Rinehart and Winston, 1970.

[10] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, 1989, pp. 233–248.

[11] H. Deng, G. Runger, and E. Tuv, "Bias of Importance Measures for Multi-valued Attributes and Solutions," *Artificial Neural Networks and Machine Learning – ICANN 2011*, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 293-300.

[12] T. Zhang, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods," *AI MAGAZINE*, vol. 22, 2001, pp. 103-104.

[13] B. Dasarathy, *Nearest neighbor (NN) norms : nn pattern classification techniques*, Los Alamitos Calif. ;Washington: IEEE Computer Society Press ;;IEEE Computer Society Press Tutorial, 1991.

[14] S.A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6, Apr. 1976, pp. 325-327.

[15] P. Hall, "Choice of neighbor order in nearest-neighbor classification," *The Annals of Statistics*, vol. 36, Oct. 2008, pp. 2135-2152.