Shape Analysis and Measurement for the HeLa cell classification of cultured cells in high throughput screening

Academic supervisor

Bjorn Olsson
School of Humanities & Informatics
University of Skövde, Sweden

External supervisors

Dr. Peter Wisskirchen and Matthias Jungmann
Fraunhofer-Institut für Angewandte Informationstechnik FIT
Sankt Augustin, Germany

ANM Enamul Huque
School of Humanities & Informatics
University of Skövde, Sweden

Shape Analysis and Measurement for the HeLa cell classification of cultured cells in high throughput screening

ANM Enamul Huque

Submitted by ANM Enamul Huque to the University of Skövde as dissertation towards the degree of M.Sc. by examination and dissertation in the School of Humanities and Informatics.

February 2006

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

_____

ANM Enamul Huque

Acknowledgements

I would like to thank my academic supervisor, Associate Professor Dr. Björn Olsson who was not only guiding me for the thesis but also for the entire duration of my Masters program. Thanks for his kindness and generosity.

I am thankful to Prof. Dr. Thomas Berlage of Fraunhofer-Institut für Angewandte Informationstechnik for giving me a chance to work with his team. My gratitude goes to Dr. Peter Wisskirchen and Matthias Jungmann for their technical guidance towards my thesis. Without their help it might have been difficult to come this far.

I also thank to Jane Synnergren and Zelmina Lubovac for their important comments on my thesis.

I would like to dedicate my thesis to my wife for being with me in hard time and for encouraging me from the beginning of the program.

Shape Analysis and Measurement for the HeLa cell classification of cultured cells in high throughput screening

ANM Enamul Huque

Abstract

Feature extraction by digital image analysis and cell classification is an important task for cell culture automation. In High Throughput Screening (HTS) where thousands of data points are generated and processed at once, features will be extracted and cells will be classified to make a decision whether the cell-culture is going on smoothly or not. The culture is restarted if a problem is detected. In this thesis project HeLa cells, which are human epithelial cancer cells, are selected for the experiment. The purpose is to classify two types of HeLa cells in culture: Cells in cleavage that are round floating cells (stressed or dead cells are also round and floating) and another is, normal growing cells that are attached to the substrate. As the number of cells in cleavage will always be smaller than the number of cells which are growing normally and attached to the substrate, the cell-count of attached cells should be higher than the round cells. There are five different HeLa cell images that are used. For each image, every single cell is obtained by image segmentation and isolation. Different mathematical features are found for each cell. The feature set for this experiment is chosen in such a way that features are robust, discriminative and have good generalisation quality for classification. Almost all the features presented in this thesis are rotation, translation and scale invariant so that they are expected to perform well in discriminating objects or cells by any classification algorithm. There are some new features added which are believed to improve the classification result. The feature set is considerably broad rather than in contrast with the restricted sets which have been used in previous work. These features are used based on a common interface so that the library can be extended and integrated into other applications. These features are fed into a machine learning algorithm called Linear Discriminant Analysis (LDA) for classification. Cells are then classified as 'Cells attached to the substrate' or Cell Class A and 'Cells in cleavage' or Cell Class B. **LDA considers features by leaving and adding shape features for increased performance.** On average there is higher than ninety five percent accuracy obtained in the classification result which is validated by visual classification.

Table of Contents

1 Introduction

Computational biology or Bioinformatics covers a variety of fields including genome sequencing, biological databases, protein structure modelling, gene expression analysis and many more. It is now becoming increasingly common to use digital image processing and digital image analysis in the field of bioinformatics. Image cytometry is the measurement of cell properties from images. Digital image analysis refers to the extraction of information from images with the aid of computers [Lindblad, 2003]. Biomedical Image Analysis deals with the research and development of image-guided surgery, shape and motion measurement, spectral analysis etc.

Techniques for digital image analysis have a long history [Ballard and Brawn 1982]. They have played a part in a number of approaches to feature extraction for cells [Dawe et al, 1994; Wied et al, 1989]. Image analysis in cell cytometry was limited to image filtration and transformation to make the objects clearer in the image for the analyser or it was simply a support for manual and visual classification. Researchers have begun to use image analysis and machine learning techniques to assist in the recognition of features associated with cells [Turner et al. 1993; Wohlberg et.al, 1993; 1995]. Interaction between sub-cellular molecules can be detected with digital image analysis by High Throughput Screening (HTS).

Biologists need well-grown cells for further experiments. If the cultured cells have aberrant growth or unwanted shapes, further experiments (for example gene expression analysis) will be misleading. So, it has to be ensured that the cell-culture is going on smoothly. For this, it is often essential to differentiate two kinds of cells in culture; cells which are in cleavage and cells which are growing normally.

HeLa cells, which are used in this study, are continuous cell lines as these can continue to divide indefinitely. HeLa cells are cancer cells and cancer results from an accumulation of mutations that activate proliferation-promoting genes (proto-oncogenes) and inactive proliferation-suppressing genes (tumour-suppressor genes) in a single cell and its progeny, which therefore proliferate without restraint [Alberts et al., 1998]. The name HeLa came from the lady named **He**nrietta **La**cks, who had cervical cancer. HeLa cells were extracted from her cervix.

Normal growing HeLa cells are attached to the substrate and floating rounded cells indicate that they are in cleavage. Their physical properties are different and they will have different mathematical properties. Presence of the correct proportion of these two kinds of cells ensures that cells are growing smoothly. In case of HeLa cells, at any time point if there are more rounded floating cells than attached cells, it indicates that there are some problems in the culture which can be in the culture medium or temperature etc. Suspending rounded cells in HeLa cell culture indicates that they are ready to divide or they were just divided. Once they have been divided they are again attached to the substrate. Dead cells also float and look round. So if there are more rounded cells floating, it means that either they are overgrown or dead cells. Either of these cases is

unwanted and this situation is to be managed in real time so that if needed the culture process can restart [Cann, 2000].

Real time processing needs automation. There may be a huge number of cells to be classified. For a human, it is obviously tiring to accomplish this job. Instead, it is preferable to have it done by a machine. Another issue will be the efficiency. A computer will be able to analyse thousands of cells and classify them more rapidly than a human. Thus, high throughput screening should be performed. HTS is obtained through a combination of modern robotics and other specialized laboratory hardware. It allows a researcher to effectively conduct hundreds of scientific experiments at once. A *screen*, in this context, is the larger experiment, with a single goal (usually testing a scientific hypothesis), to which all this data may subsequently be applied [Hann and Oprea, 2004].

From the above discussion it is clear that three different things should be accomplished, and they are feature extraction, classification and automation. The scope of this thesis is feature extraction and classification of cancer cells. There are as many as twenty features that are found from mathematical properties of objects and from image processing techniques. Those morphological properties are described in the method section. It is intuitively anticipated that feature values will be different for two different types of cells. For example, a cell in cleavage will be more round and thus the roundness feature will discriminate an attached cell which is irregular in shape, from a floating cell in cleavage, which is regular in shape.

The input data here is raw images from cells. There may be thousands of cells in a single image. They need to be separated from one another. This is done by segmentation. Segmentation is a process in which an image is subdivided into its parts or objects, or, more simply stated, it is a process to isolate objects from background [Gonzalez and Woods, 2003]. There are many techniques for segmentation but a manual process is used here for the experiment, since the topic automatic segmentation is beyond the scope of this thesis. When a single cell is found by segmentation and isolation, its different features are calculated by digital image analysis. These techniques are repeated for all cells in the whole image or a part of the image.

When the features are found for all the cells, they are fed as input data to a machine learning algorithm called Linear Discriminant Analysis (LDA) for classification. LDA is a classical statistical approach for classifying samples of unknown classes, based on training samples with known classes. LDA constructs a line between the training data of the two sets in a way that some optimization criterion is fulfilled (maximal distance of means versus minimal scatter). Then the classification process is defined by selecting this line as a separating criterion although the training set is, in general, not perfectly separated. This classification algorithm classifies the given cells as irregular attached cells and floating spherical cells which here are called Cell Class A and cell Class B.

Automation is to be carried out in all the steps except input / output of the cell culture (flasks). So the above processes (feature selection, segmentation and classification) are all to be automated. In addition, in a fully automated cultivation system optical

monitoring, cell handling and system control are automated. This work is a part of a project called Live Cell Monitoring: Development of a System for Cultivation and Monitoring of Living Cells. The aim of this thesis is to design and implement a fully automated feature selection and classification algorithm. Other parts of the Live Cell Monitoring project, such as cell handling, automatics image segmentation and system control are beyond the scope of the thesis.

Chapter 2, Background, presents biological and image processing background material relevant to this work. Chapter 3, Project describes the entire project in general which is the alliance of four different Fraunhofer institutes and the contribution of this thesis work to the project. Chapter 4, Methods, describes different mathematical and image features which were used to classify the cells. It also describes the classification algorithm- LDA in brief. Chapter 5, Results and Analysis, presents classification results of the experiment on five different HeLa cell images using the features described in Chapter 4, Methods. Chapter 6, Discussion and conclusion, discusses about the prediction capabilities of the classification algorithm using the feature set and concludes with topics of misclassification and computational overhead. Chapter 7, Future work, describes about enhancement of this work in the area of automatic image segmentation, feature inclusion and overhead reduction which were beyond the scope of this thesis.

2 Background

2.1 Biological Background

Cell classification is an important assay in cell culture. Once in culture, cells exhibit a wide range of behaviours, characteristics and shapes. Cultured cells are usually described based on their morphology, that is shape and appearance, or their functional characteristics. There are three types of cells: Epithelial, Fibroblast and Lymphoblast-like cells. Each type of cell has some different properties.

Epithelial cells stay attached to the culture plate. They appear flat and are normally polygonal. Fibroblast cells are also attached to the plate or substrate but appear elongated and bipolar. They form swirls in culture. Lymphoblast cells do not stay attached to the substrate but remain suspended and are round. These are shapes a cell may have but culture condition plays an important role in determining cell shape [Ryan, 2003].

Cell shape is one of the important characteristics which help to evaluate the general health of cells in culture. Determining the general health of the cell culture is important as they will be subsequently experimented for further research such as cell division, programmed cell death, protein identification and gene expression etc. Cell culture needs to be constantly monitored; otherwise cells will deviate from normal structure and will have unwanted characteristics such as changed expression profile that can adversely affect the experiments.

Since both normal cells and cancer cells can be grown in culture, the basic differences between them can be studied closely. In addition, it is possible, by the use of chemicals, viruses and radiation, to convert normal cultured cells to cancer-causing cells. Thus, the mechanisms that cause the change can be studied. Cultured cancer cells also serve as a test system to determine suitable drugs and methods for selectively destroying some types of cancer [Ryan, 2003].

Figure 2.1 shows healthy HeLa cells in culture. Cells in this figure have defined outer membranes. They are angular shaped and the growth pattern forms a patchy monolayer. In the figure, few rounded cells can be found, which is normal for actively dividing population.

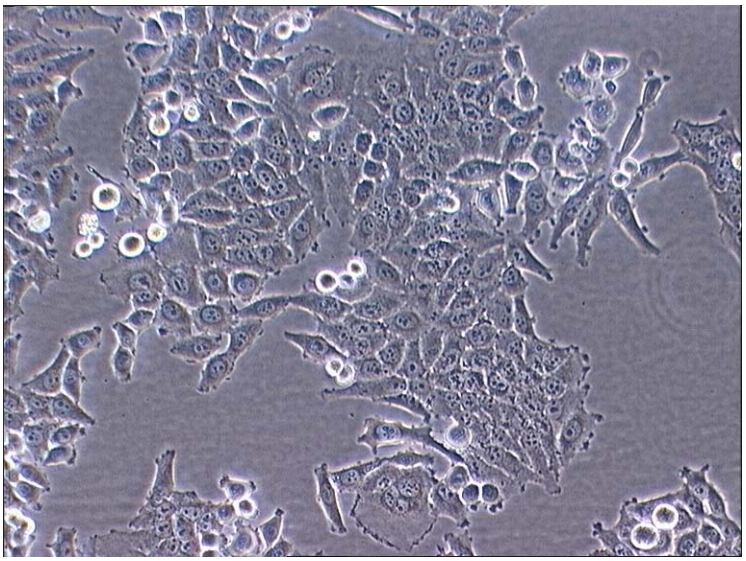


Figure 2.1 healthy HeLa cells

But too many stressed and rounded cells as in figure 2.2 indicate a problem in cell culture condition which can lead to excessive apoptosis (programmed cell death) or necrosis (swell and breakdown). In this figure the cells are overly crowded. The cells are stressed by the culture condition. In this situation there will be morphological changes and increase in the amount of sub-cellular granular particles in the overly crowded culture. Experiment with such cells will be affected as their expression profile is changed.

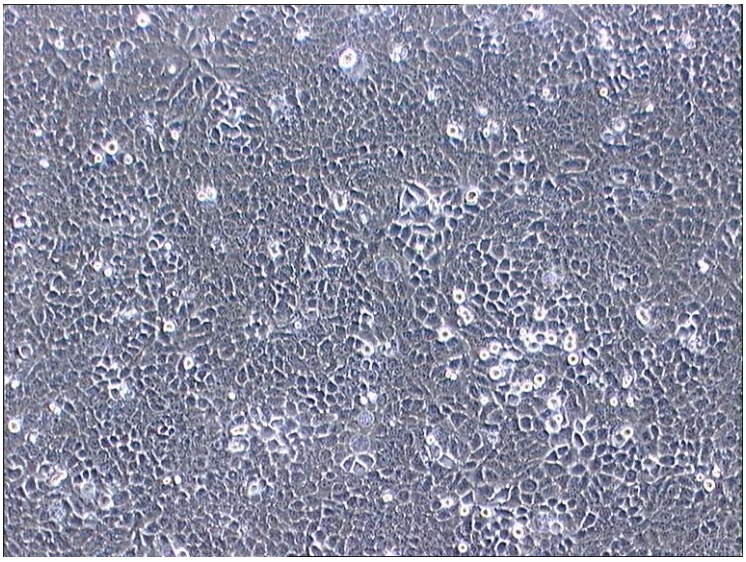


Figure 2.2 stressed HeLa cells

In good culture condition immortalised cells like a HeLa cell-line will grow indefinitely. To obtain best experimental results, it is recommended that the cells not to passage more than sixty times even in good culture condition. This has to be taken care of else cells will have distorted or changed gene expression profiles.

As described above, the number of cells in cleavage will be fewer than the number of normal growing attached cells. The Live Cell Monitoring system counts the number of cells of both types and checks the proportion. If the proportion between the cell counts is correct, the culture process is carried on. Otherwise, the condition is reported and a new culture is started.

2.2 Image Processing steps for supervised Cell Classification

Before the description of chapter 2.3, Supervised learning and training data, a short overview of the main image processing steps are presented as required in supervised cell classification.

2.2.1 Digital Image and Analysis

The term image refers to a two-dimensional light intensity function $f(x, y)$, where x and y denote spatial coordinates and the value of f at any point (x, y) is proportional to the brightness or gray level. A digital image is an image $f(x, y)$ that has been discretized both in spatial coordinates and brightness. A digital image can be considered a matrix where each combination of row and column coordinates identifies a point in the image and the corresponding matrix element value identifies the gray level at that point. The elements of such a digital array are called image elements or pixels [Gonzalez and Woods, 2003].

Cell analysis by digital image processing is the main focus of this work. Image analysis techniques are being used in the area of cytometry. Initially most of these applications were focused on performing image transformations to make an image clearer and brighter to a human analyzer [Dawe et al, 1994; Wied and Bartels, 1989; Wittekind and Schulte, 1987]. Feature extraction from cell images, cell migration etc are very recent applications of digital image cytometry [Wohlberg et al, 1995]. In this work, image analysis is used specifically to extract descriptive features of HeLa cells growing in culture.

2.2.2 Image acquisition

Cell images are obtained by assembling a camera to the top of the microscope. The camera takes the image of the culture plate below the microscope and sends the image to the attached computer in the common image formats, such as TIFF, for example. Then, for further analysis, the coloured image is transformed to greyscale which has pixel values in the range 0 to 255.

2.2.3 Cell segmentation

One single cell is analyzed at a time. Hence, the cell to be analyzed needs to be isolated from the rest of the group before any further analysis can be carried out. Finding the outline of a cell is an important but difficult task. Difficult in the sense that microscopic images are noisy. Sometimes it is difficult to differentiate between background and foreground. Cells are in foreground and those have to be isolated in such a way that they are not associated with the background. For that reason image has to be pre-processed before segmentation. Pre-processing consists of smoothing of the image to reduce the effect of noise or filtering to reduce the effect of smoothing. Since a general solution for

this problem is not available, different segmentation algorithms are used. Here, manual segmentation is used for experiment, as this paper is concerned with classification rather than with segmentation.

2.2.4 Feature extraction

The next step in the analysis process is to extract descriptive feature measures from the segmented cells. There are many different, more or less general, features described in the literature [Rodenacker and Bengtsson, 2003]. However, the chosen features, if to be of any use in the further analysis, have to reflect the property of interest. When working on a real world application, it is rarely enough to only use general purpose features. To achieve good results from the data analysis, it is almost always fruitful to measure additional features specifically designed to capture the property of interest [Lindbald, 2003]. Features should also have some important properties. Absolute location or positioning (horizontal or vertical) of cells are irrelevant to classification. So the features should be invariant to translation. Rotation is also irrelevant and thus features should be invariant to rotation. Also the size of the cell may not be important. So the shape features should be invariant to scale [Duda et al., 2001]. In the Method section a large number of features are described which are implemented for the experiment.

2.3 Supervised learning and training data

Supervised learning is generally more efficient, and therefore to be preferred in those cases where it is possible to apply. And that in this case, it is possible to apply it because we know which classes should be learnt and we have lots of examples from each class which we can use in the training process. In supervised learning training data should specify what one is trying to learn (the class). On the other hand, in unsupervised learning training data does not say what one is trying to learn [Manning and Schütze, 1999].

The features are input to a learning algorithm of type supervised learning named Linear Discriminant Analysis (LDA). This algorithm classifies the observed cell as attached or floating depending on the features trained on. Supervised learning is a machine learning technique for classification from training data. In supervised learning a class is predicted for the data it is trained on. In training set, data objects (cells) and output (label) should be present. After seeing a small number of training data, supervised learning algorithm should predict the value (label) of the function on input object or new pattern.

The input to the classification algorithm of cells should be the basic feature of an image object that is pixels intensities. But pixels-based features suffer form dimentionality problem. In the subsections below motivation and importance of concrete features which are combination of pixels are presented.

2.3.1 Supervised cell classification

In the case where cell should be classified, the user can be seen as a human supervisor who tells the computer system how to classify cells into different classes. For this purpose, given cell centers (detected by a preprocessing module which is not part of this

thesis) are identified by mouse clicks together with the attachment of a label such as Class A and Class B.

2.3.2 Training sets

As a result of the above training step, different sets of image positions (cell centers) are given together with the label of the respective class. The basic information relevant for the classification consists of a neighbourhood of cell centers consisting of about 30×30 pixels. This neighbourhood can be interpreted as a set of about 1000 basic features consisting of image intensities or the values [0, 1] for the case of a binary image.

In principle, a learning algorithm such as LDA, mainly used in this work, can be based on this feature set. In practice, however, pixels as features are not useful for this task mainly for two reasons. First, single pixels are corrupted by noise and will not describe the cells as precise as is needed by the learning algorithm. Secondly, pixels of a cell neighbourhood can be interpreted as a very high (about 1000) dimensional feature vector. This means that the dimension of the feature space is much higher than the size of the training set. In this case, nearly all learning algorithms, in particular LDA, are able to discriminate the samples of the training set perfectly. But this discrimination has very poor generalization quality.

To solve this dilemma, quite sophisticated combinations of pixels that result in a set of few features are required which are much more robust against noise and have a much better generalization quality. Because these features are based on the combination of neighbourhood pixels, they can be seen as a result of feature reduction. These features are the main topic of this thesis. These features can be roughly separated into two groups. One group consists of features directly operating on pixels by combining them, for example features derived by moment analysis. The other group consists of features working on boundaries of cells, delivered by a segmentation step. Figure 2.3 shows the principle in a schematic view. In figure 2.3 (a), (a two dimensional) projection of well separated training sets is shown with bad generalisation. In figure 2.3 (b), two training sets based on shape features are shown with (possibly) bad separating quality on training sets but good generalisation quality are displayed.

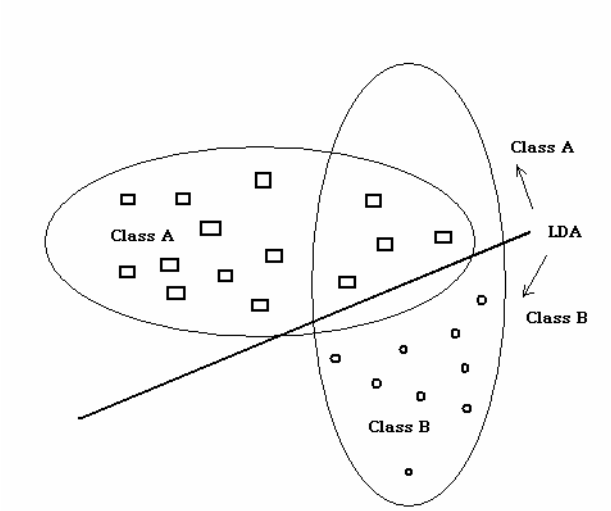


Fig 2.3 (a) Pixel based generalisation

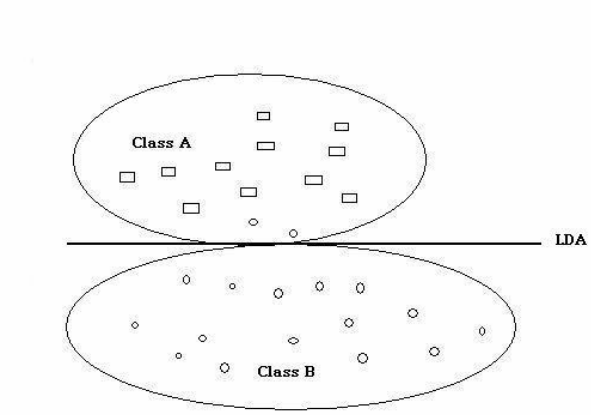


Fig 2.3 (b) Feature based generalisation

2.4 Related work

In the paper 'A feature set for cytometry on digitized microscopic images' [Rodenacker and Bengtsson, 2003], a possible feature set was acquired. The feature set described is divided into morphometric, densitometric, textural and structural features. The main goal of the paper was to bring attention to the need of a common and well defined description of features used in cytometry and histological studies. This paper in general is a collection of features with definitions.

In the paper 'Development of algorithms for digital image cytometry' [Lindbald, 2003], appropriate algorithms were presented for image segmentation for fluorescence microscope images of cultured cells. It focuses on the development and compilation of robust image analysis tools, enabling quantitative measurements of various properties of cells and cell structures. An effort was also made in the areas of feature extraction and statistical data analysis. A classification method that separates individual cells into three classes, depending on their level of activation, is described. The method is based on analysis of time series of images. The major contribution of this paper was automatic segmentation of nucleus and cytoplasm. The feature set is relatively small and not the main focus of the work although some specific features were measured. The set of features was used to classify cells depending on level of activation.

The paper 'Creating classification features for biological images' [Naik, 1998], presents a system based on image processing and machine learning techniques to characterize cellular events occurring during the process of cell division, meiosis, and to classify images of cells exhibiting these events. The system is based on extraction of features from cell images and construction of a classifier that distinguishes cell images of one type from other.

The paper ‘Classification of cultured mammalian cells by shape analysis and pattern recognition’ [Olson et al., 1980] presents a method for classifying cultured cells on the basis of shape characteristics. The authors used hierarchical cluster analysis and nearest neighbour analysis for classification of cells. LDA was also used but it provided only a slight improvement. 4 to 5% misassignments were obtained when they used twenty descriptors. The classification result was almost same when the authors used 8 features instead of 20.

The novelty of the work presented in this thesis lies in the application of the work. The features are implemented in such a way that these can be considered as a broad and complete library in Java, based on a common interface. This library can be easily extended and integrated into another application.

In this work, the emphasis is on feature selection for all three types of cultured cells, i.e. epithelial, lymphoblast, fibroblast-like cell lines. For the experiment, HeLa cells are chosen to be classified by the feature set. For every feature, a clear definition is given so that the experiment may be repeated.

In some papers the authors used LDA for classifying cells mainly based on a restricted feature set but in this work a broad range of features were used. New features may be added without degrading the classification performance because the classification algorithm retains the best discriminatory features by leaving and adding shape features.

There are some new features used in this work. These features were not mentioned in any other cell classification papers. Those features are orientation, eccentricity by moment, Euler number, spread of the object, elliptic variance, circular variance, sphericity, solidity and bending energy.

3 Project

The name of the project in which this thesis work is included is called **Live Cell Monitoring: Development of a System for Cultivation and Monitoring of Living Cells.**

3.1 The alliance

In the frame of "Live Cell Monitoring – non-invasive analysis, quality assurance and process control of the cultivation and differentiation of (stem) cells" four institutes of the Fraunhofer Gesellschaft have established an alliance in order to develop an automated system for cell cultivation. The reproducibility and comparability of cell cultures presently suffers from different and subjective handling by the technical personnel in the cell culture laboratory. For this reason the institutes bundled their competencies in optics, automation, informatics and cell systems to develop a platform technology for reproducible and standardized cell cultivation as shown in figure 3.1.

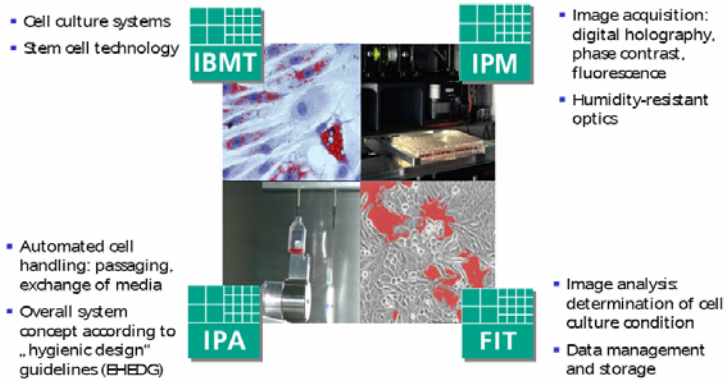


Figure 3.1 the alliance of four institutes

3.2 The system concept

The state-of-the-art of cell cultivation can be described on the one hand as manual cultivation of cells in a cell culture laboratory containing an incubator for the climate control of the cell culture and hand operated microscopes for the daily glance to decide whether passaging or media exchange is necessary. On the other hand there are systems on the market that carry out the cell cultivation a rigid industrial process without any inspection of the cell culture. The aim of the Fraunhofer alliance is to develop an automatic cell cultivation system where after the input of the cell culture flask the whole cell culture procedure is executed including optical monitoring, addition of factors, exchange of media and passaging. This is only feasible by the optical image acquisition of the cell cultures and user-friendly image analysis software that determines the cell culture condition in order to control the cell culture process. Thereby the cell culture process can be documented and archived, too. Finally the expanded cell line (contained in flasks or micro well plates) can be taken out and applied for cell-based screening, toxicity tests or other purposes. The whole cultivation process takes place under cultivation

conditions (37 °C, CO2, 95 % humidity) to leave the cell culture in an optimal environment and to avoid stress because of climate changes.

3.3 Contribution of this work to the project

The aim of this project is to produce a good descriptor-set that will classify any cell type in an experiment. For this work, HeLa cells are used. Two dimensional binary HeLa cells are experimented as objects. Several object features which are generally used in biomedical image analysis are studied and presented with their mathematical description and implemented in Java. The issue here was to see the growth of the cancer cells, so the time-varying cancer cell (HeLa) images are obtained for classification (by those descriptors or features) such as cells on division and cells on normal growth. So, this classification problem is binary. Thus LDA is well suited candidate for this kind of classification. LDA was implemented with a special capability of leaving and adding object features for increased performance for classification.

This work goes under the Image analysis: determination of cell culture condition of FIT as a subset of the whole project (Fig. 3.1).

4 Method

A good set of descriptor features should include the features that capture the most important properties of an object and can be used to identify the object uniquely. An object can be identified by its two or three dimensional geometrical properties. Such properties could be area, perimeter, or moments. An appropriately selected set of such features carries sufficient information for the identification of an object and thus for an individual cell. Some features which are not directly geometrical but based on geometry can be seen as hybrid features. When geometrical features are used, a relatively small feature vector is enough for describing an object and thus results in significant data compression. Features are extracted from representations of shapes like boundary chain code or from a binary image [Pitas, 2000]. Features extracted for this project to describe a two dimensional HeLa cell (object) are (in almost all cases) scale, rotation and translation invariant, as mentioned before, and they are presented in the following subsections with their definition.

4.1 Perimeter and convex perimeter

Perimeter is an important feature of an object. Contour based features which ignore the interior of a shape, depend on finding the perimeter or boundary points of the object [Celebi and Aslandogan, 2005]. The perimeter of an object is given by the integral as follows:

$$T = \int \sqrt{x^2(t) + y^2(t)}dt \qquad \text{(Eq.1)}$$

This perimeter is used for the parametric boundary representation. By the aid of a boundary following algorithm, the object perimeter can be found out. If $x_1,.....x_n$ is a boundary coordinate list, the object perimeter is given by:

$$T = \sum_{i=1}^{N-1} d_i = \sum_{i=1}^{N-1} \left| x_i - x_{i+1} \right| \qquad \text{(Eq.2)}$$

The convex perimeter is defined by the convex hull of an object. The perimeter of the convex hull that encloses the object is the convex perimeter as shown in Fig. 4.1, redrawn from [Wirth, 2001].

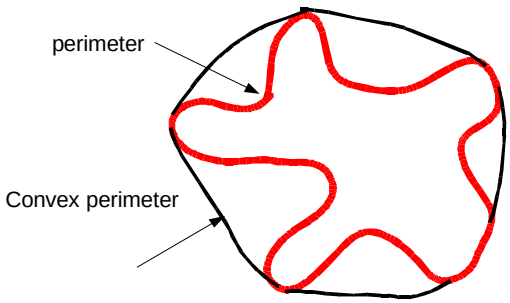


Figure 4.1 object perimeter and convex perimeter

4.2 Moment Descriptors

Moment-based shape descriptors are used when a region-based analysis of the object is performed. Region-based analysis exploits both boundary and interior pixels of an object. These shape descriptors are more robust to noise and distortions. Moment is popular for region-based analysis as central moments are invariant to translation, rotation and scale. They are also computationally simple [Celebi and Aslandogan, 2005]. Moment analysis describes essential and frequently used shape features. For a continuous image $f(x, y)$, the moment is given by:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)dxdy \quad p,q = 0,1,2... \qquad \text{(Eq.3)}$$

For example, the center of gravity is a very common feature. The coordinates of the center of mass is given by the object moments as follows:

$$\bar{x} = \frac{m_{10}}{m_{00}} \qquad \bar{y} = \frac{m_{01}}{m_{00}} \qquad \text{(Eq.4)}$$

The central moment is defined by using the center of mass:

$$\mu_{pq} = \int_{\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q f(x,y)dxdy \quad p,q = 0,1,2... \qquad \text{(Eq.5)}$$

Normally the image is discrete and for that the moment and central moment are defined by:

$$m_{pq} = \sum_i \sum_j i^p j^q f(i,j) \qquad \text{(Eq.6)}$$

$$\mu_{pq} = \sum_i \sum_j (i-\bar{x})^p (j-\bar{y})^q f(i,j) \qquad \text{(Eq.7)}$$

Indices i, j corresponds to the x and y respectively. Images in this work are binary and for a binary image $f(i, j)$, the moment calculations are given as follows:

$$m_{pq} = \sum_i \sum_j i^p j^q \qquad \text{(Eq.8)}$$

$$\mu_{pq} = \sum_i \sum_j (i-\bar{x})^p (j-\bar{y})^q \qquad \text{(Eq.9)}$$

The object area of a binary image is given by its moment m_{00}. The following relationship gives central moments by up to third order moments [Pitas, 2000]:

$$\mu_{00} = m_{00} = \mu \qquad \text{(Eq.10)}$$

$$\mu_{10} = \mu_{01} = 0 \quad \text{(Eq.11)}$$

$$\mu_{20} = m_{20} - \mu\bar{x}^2 \quad \text{(Eq.12)}$$

$$\mu_{11} = m_{11} = \mu\bar{x}\bar{y} \quad \text{(Eq.13)}$$

$$\mu_{02} = m_{02} - \mu\bar{y}^2 \quad \text{(Eq.14)}$$

$$\mu_{30} = m_{30} - 3m_{20}\bar{x} + 2\mu\bar{x}^3 \quad \text{(Eq.15)}$$

$$\mu_{21} = m_{21} - m_{20}\bar{y} - 2m_{11}\bar{x} + 2\mu\bar{x}^2\bar{y} \quad \text{(Eq.16)}$$

$$\mu_{12} = m_{12} - m_{02}\bar{x} - 2m_{11}\bar{y} + 2\mu\bar{x}\bar{y}^2 \quad \text{(Eq.17)}$$

$$\mu_{03} = m_{03} - 3m_{02}\bar{y} + 2\mu\bar{y}^3 \quad \text{(Eq.18)}$$

The angle between the major axis of the object and axis x is called the object orientation and is given by the angle θ as follows:

$$\theta = \frac{1}{2}\arctan\left(\frac{2\mu 1 1}{\mu_{20} - \mu_{02}}\right) \quad \text{(Eq.19)}$$

Eccentricity of the object is calculated by [Levine, 1985]:

$$\varepsilon = \left[\frac{\mu_{02}\cos^2\theta + \mu_{20}\sin^2\theta - \mu_{11}\sin 2\theta}{\mu_{02}\sin^2\theta + \mu_{20}\cos^2\theta + \mu_{11}\cos 2\theta}\right]^{1/2} \quad \text{(Eq.20)}$$

Following is an alternative definition of eccentricity [Jain, 1989]:

$$\varepsilon = \frac{\left(\mu_{02} - \mu_{20}\right)^2 + 4\mu_{11}}{A} \quad \text{(Eq.21)}$$

In the above equation, A is the object area. By second order central moment the object spread is defined as follows [Hu, 1962]:

$$S = \mu_{02} + \mu_{20} \quad \text{(Eq.22)}$$

4.3 Major and Minor axes

Major and minor axes are the simplest of all features but yet important. They give essential information of an object such as elongation, eccentricity etc. They are also used to find other features like elliptic variance.

The major axis points are the two points in an object where the object is more elongated and where the straight line drawn between these two points is the longest [Costa and Cesar, 2001]. Major axis points are calculated by all possible combinations of perimeter

pixels where the line is the longest as shown in Fig 4.2, redrawn from [Wirth, 2001]. The length of the major axis is given by:

$$\text{Major-axis length} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad \text{(Eq.23)}$$

where (x_1, y_1) and (x_2, y_2) are the coordinates of the two end points of the major axis.

The minor axis is drawn perpendicular to the major axis where this line has the maximum length. Once the end points of the minor axis have been found, its length is given by the same equation as the major axis length. It is also called the object width.

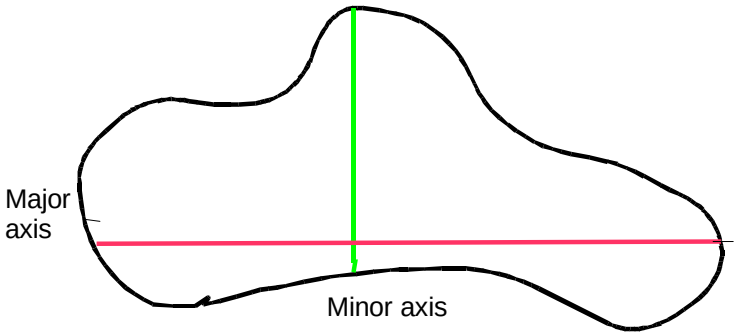


Figure 4.2 major and minor axis

4.4 Compactness

Compactness is the ratio of the area of the object to the area of a circle. The circle is defined by the same center of mass as the object and its radius is defined by the average distance from the center of mass to the perimeter of the object [Archard et al, 2000]:

$$\text{Compactness} = \frac{4\pi \cdot area}{(perimeter)^2} \qquad \text{(Eq.24)}$$

As a circle is the object with the most compact shape, a circle takes the maximum value of compactness, that is 1, while a square has the value of $\pi/4$. Fig 4.3, redrawn from [Wirth, 2001], shows different compactness values for different objects.

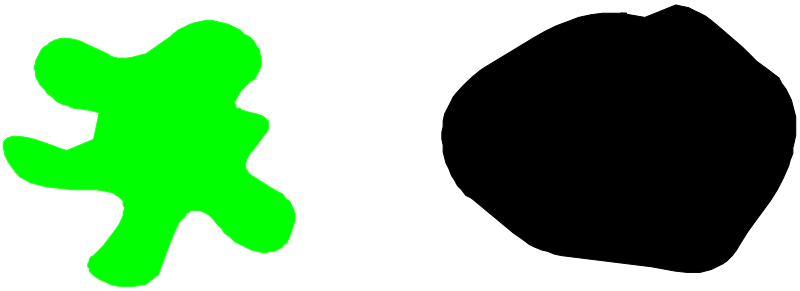


Figure 4.3 different compactness of objects

This feature is motivated by the fact that attached normal growing cells will have a low compactness value compared to that of floating cells in cleavage.

4.5 Elongation

Elongation [Jenkin, 1997] is defined to be the ratio between the width and length of the minimum bounding box as shown in Fig 4.4, redrawn from [Wirth, 2001]:

$$\text{Elongation} = \frac{width_{bounding-box}}{length_{bounding-box}} \qquad \text{(Eq.25)}$$

The result varies from 0 to 1. If the object or cell is more or less like a square or circle, the values get closer to 1 and if the objects deviate from the above objects then it gets closer to 0.

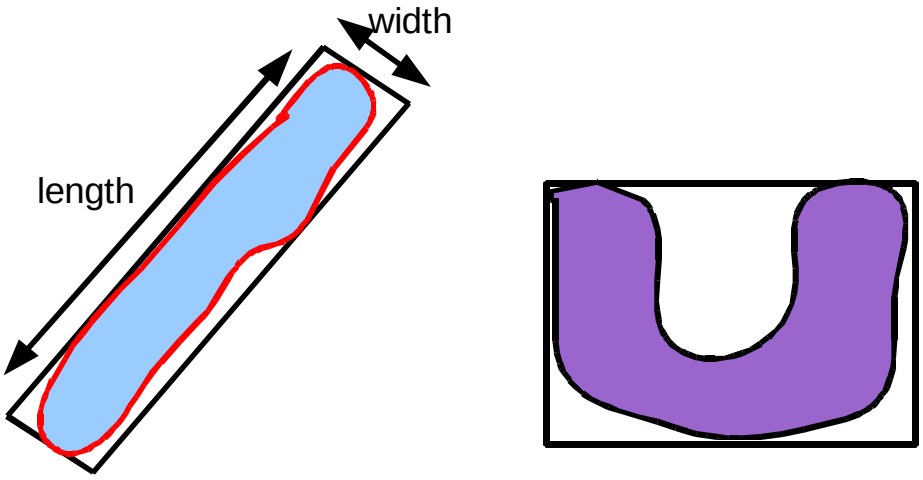


Figure 4.4 object on the left is more elongated than the object on the right

4.6 Eccentricity

Eccentricity is the ratio between the length of the short axis to the long axis [Gonzalez and Woods, 2003] as defined in the following equation.

$$\text{Eccentricity} = \frac{axislength_{short}}{axislength_{long}} \qquad \text{(Eq.26)}$$

The value of eccentricity is between 0 and 1. Eccentricity is also called ellipticity with respect to minor axis and major axis of the ellipse. If the major axis gets longer, eccentricity gets higher (by the alternative definition of eccentricity Eq. 21) as shown in Fig 4.5, redrawn from [Wirth, 2001].

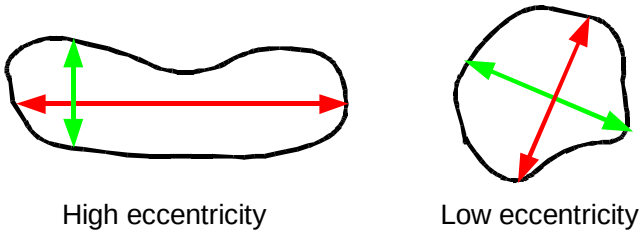


Figure 4.5 Object eccentricity

Irregular cells are longer and expected to have higher eccentricity than that of round floating cells.

4.7 Circularity or Roundness

Area-to-perimeter ratio is the measure of roundness or circularity [Castleman, 1996]. But local irregularities are not reflected by this feature. It is defined as:

$$Roundness = \frac{4\pi \cdot area}{(convexPerimeter)^2} \qquad \text{(Eq.27)}$$

A circle gets the value of 1, while objects with bumpy boundaries get lower values as shown in Fig 4.6, redrawn from [Wirth, 2001].

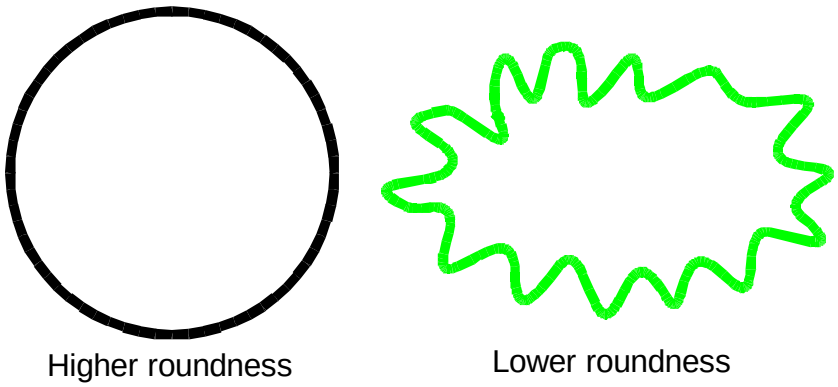


Figure 4.6 roundness of object

The circularity or roundness feature is likely to be one of the important features because it should contribute to classify two different types of cells as they will have different values. Cells in cleavage are normally round, so their roundness value will be higher and on the other hand normal growing cells are irregular so their roundness value will be lower.

4.8 Sphericity

Sphericity [Ya, 2003] is the ratio between the radii of the inner circle and the outer circle of the object as shown in Fig 4.7, redrawn from [Wirth, 2001]. For a circle the value reaches 1.

$$spherity = \frac{R_{inscribing}}{R_{circumscribing}} \qquad \text{(Eq.28)}$$

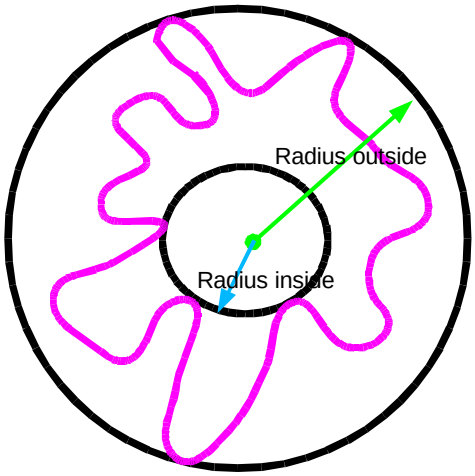


Figure 4.7 spherity defined by Radius inside and Radius outside

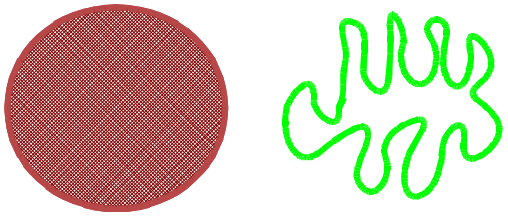
As seen in figure 3.8 a circular cell will be more spherical than a cell with an irregular boundary.

4.9 Convexity

The measure of convexity of an object is the ratio of the perimeter of the convex object to the original perimeter of the object. It is a relative measure of how much an object differs from a corresponding convex object [Shipley and Kellman, 2001]:

$$convexity = \frac{convexPerimeter}{perimeter} \qquad \text{(Eq.29)}$$

The value of convexity is 1 for a convex object and the value is lower when the perimeter (of the object) is rough as shown in Fig 4.8, redrawn from [Wirth, 2001]. Thus an irregular shaped cell will have a low convexity value, while a round (or closer to round) cell will have a high convexity value.



High convexity Low convexity

Figure 4.8 different convexity of objects

4.10 Aspect Ratio

Aspect Ratio is the ratio between height and width of the corresponding object. By this feature, slender (long and thin) object can be separated from circular or square object [Castleman, 1996].

$$aspectRatio = \frac{height}{width} \qquad \text{(Eq.30)}$$

Compared to elongation and eccentricity, aspect ratio is not rotation invariant and is just describing height versus width. This simple feature may not be useful for cell classification, but is added for completeness.

4.11 Solidity

In simple terms density is mass per unit volume [Derry, 2002]. But in two dimensional image objects this can be defined as the ratio between the area and convex area of the same object:

$$solidity = \frac{area}{convexArea} \qquad \text{(Eq.31)}$$

For a solid object or cell, this value is 1, while the value is lower for an object or cell having a rough perimeter or an object which has holes in it as shown in Fig 4.9, redrawn from [Wirth, 2001].

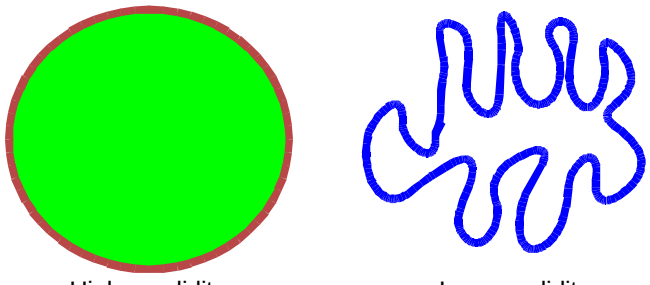


Figure 4.9 solidity of different objects

4.12 Shape variances

Two different shape variances are presented and they are circular variance and elliptic variance. As spherical cells will have less variance (error) compared to a circle or an ellipse than cells with irregular boundaries, these properties should be good discriminative factors.

When the center of mass of an object and a (x, y) point of ellipse form a line with a specific angle, there should be a single point of the object perimeter which lies in the same line and angle. The difference between object perimeter point and the (x, y) point of the reference object (in the same line and angle) gives the relative error, and measure of all errors is the elliptic variance. For a circle, first the mean radius is calculated from the center of mass to the border points of the object, and then the differences of the radii to the mean radius are used to calculate the circular variance.

4.12.1 Circular variance

A cell or object should sometimes be compared to a reference object, like for example a circle as shown in Fig 4.10, redrawn from [Wirth, 2001]. It is necessary to find out the variance of the object from a circle. This will give a relative measure of whether the object is round (or close to round) or not. The object is compared to a solid circle and for a perfect round object the value of the variance is 0. It means that the proportional mean squared error is 0. The value increases when the shape is complex or elongated in the major axis [Lee et al., 2003].

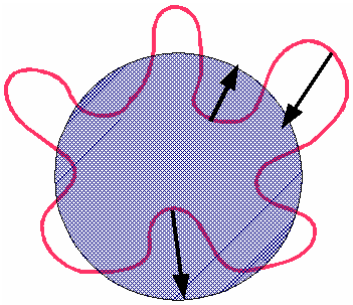


Figure 4.10 circular variance

4.12.2 Elliptic variance

Elliptic variance is similar to circular variance. A proportional mean squares error with respect to a solid ellipse is defined [Agouris and Stefanidis, 2000]. The equivalent ellipse is defined as an ellipse that has the same center of mass as the corresponding object (Fig. 3.13). The elliptical area and perimeter are derived from the major and minor axis of the equivalent ellipse. The parametric equation of an ellipse is given by:

$$x = a\cos\theta \qquad \text{(Eq.32)}$$

$$y = b\sin\theta \qquad \text{(Eq.33)}$$

where a and b are semi major and semi minor axis respectively as shown in Fig 4.11.

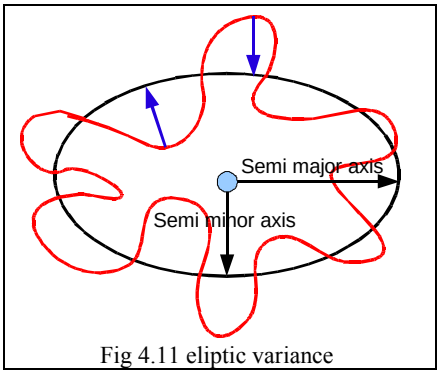


Fig 4.11 eliptic variance

4.13 Bounding Box

The bounding box is defined by the smallest rectangle [Costa and Cesar, 2000] which encloses the object as shown in Fig 4.12, redrawn from [Wirth, 2001]. The minimum area of such a bounding box is given by:

$$area = majorAxisLength * minorAxisLength \qquad \text{(Eq.34)}$$

It gives the minimum area of the box.

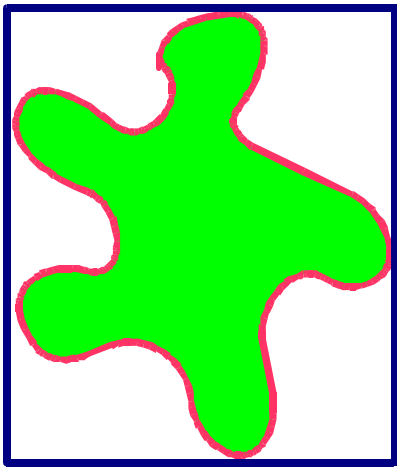


Figure 4.12 bounding box

4.14 Topological Descriptors

One way of obtaining useful global information about an object is to use topological descriptors. A topological descriptor gives information about the regions of the image plane of an object [Gonzalez and Woods, 2003]. It is unaffected by any deformation such as stretching, rotation or transformation. Connected components and holes are important

topological features and they are found out by the Euler number. The Euler number (E) is defined by the number of connected components (C) and holes (H):

$$E = C - H \qquad \text{(Eq.35)}$$

It is an important topological descriptor. This simple topological feature as said before is invariant to translation, rotation and scaling. For example the object in the Fig. 4.13 has the Euler number 0 as it has one connected component and one hole.

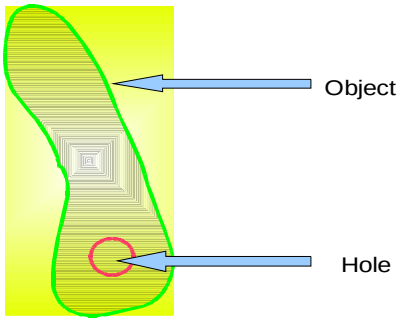


Figure 4.13 Euler number defined by number of connected components

This feature does not contribute to classification but it can be used for object filtering. A cell should not have holes in it, since it should be one single connected component. But if it contains one, it means that the image of the cell is corrupted and should be filtered out. This feature can be used for that purpose.

4.15 Boundary Descriptors

There are many features that depend on boundary descriptors of objects such as bending energy, curvature etc. For an irregularly shaped object, the boundary direction is a better representation although it is not directly used for shape descriptors like centroid, orientation, area etc [Kim et al., 2002].

Consecutive points on the boundary of a shape give relative position or direction. A 4- or 8-connected chain code is used to represent the boundary of an object by a connected sequence of straight line segments [Gonzalez and Woods, 2003]. 8 connected number schemes are used to represent the direction in this case. It starts with a beginning location and a list of numbers representing directions such as $d_1, d_2, \cdots, d_N$. Each direction provides a compact representation of all the information in a boundary. The directions also represent the slope of the boundary. In Fig. 4.14, redrawn from [Wirth, 2001], an 8 connectivity chain code is displayed where the boundary description for the boxes with red arrows will be 2-1-0-7-7-0-1-1.

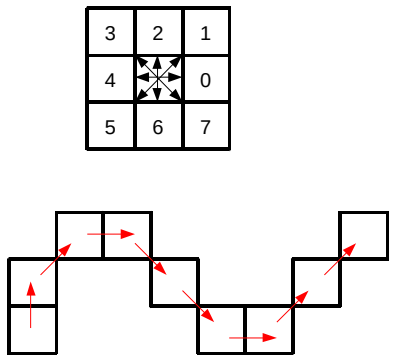


Figure 4.14 boundary descriptor

4.15.1 Curvature

The rate of change of a slope is called the curvature. As the digital boundary is generally jagged, getting a true measure of curvature is difficult. The curvature at a single point in the boundary can be defined by its adjacent line segments. The difference between slopes of two adjacent (straight) line segments is a good measure of the curvature at that point of intersection [Gonzalez and Woods, 2003].

The curvature of the boundary at (x_i, y_i) can be estimated from the change in the slope is given by:

$$\left\{\tan^{-1}\left(\frac{y_{i+k}-y_i}{x_{i+k}-x_i}\right)-\tan^{-1}\left(\frac{y_i-y_{i-k}}{x_i-x_{i-k}}\right)\right\}\ (\mathrm{mod}\,2\pi) \qquad \text{(Eq.36)}$$

Curvature (κ) is a local attribute of a shape. The object boundary is traversed clockwise for finding the curvature. A vertex point is in a convex segment when the change of slope at that point is positive; otherwise that point is in a concave segment if there is a negative change in slope as shown in Fig 4.15, redrawn from [Wirth, 2001].

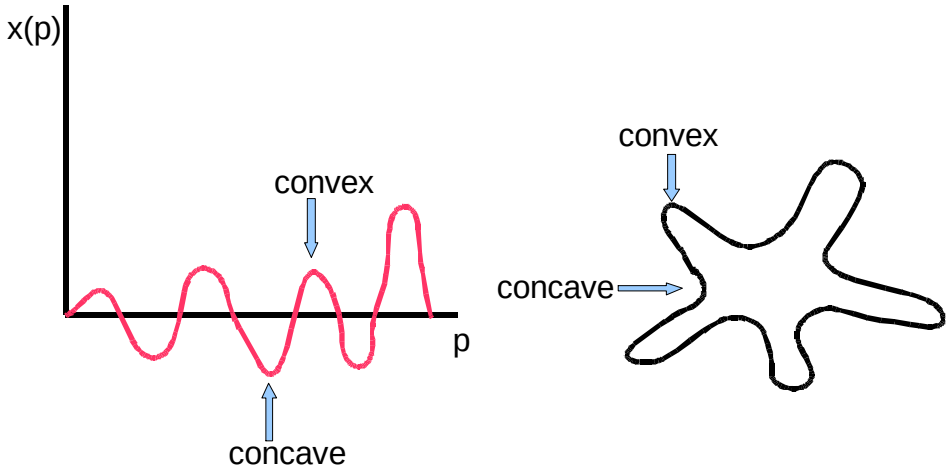


Figure 4.15 curvature of a boundary

4.15.2 Bending Energy

The descriptor called bending energy is obtained by integrating the squared curvature $\kappa(p)$ through the boundary length L. It s a robust shape descriptor and can be used for matching shapes [Costa and Cesar, 2000].

$$E_c = \frac{1}{L}\sum_{p=1}^{L}\kappa(\rho)^2 \qquad \frac{2\pi}{R} \le E_c \le \infty \qquad \text{(Eq.37)}$$

The value $2\pi / R$ will be obtained as its minimum for a perfect circle with radius R and the value will be higher for an irregular object.

4.15.3 Total Absolute Curvature

Total absolute curvature is the curvatures added along the boundary points and divided by the boundary length.

$$\kappa_{total} = \frac{1}{L}\sum_{p=1}^{L}\left|\kappa(\rho)\right| \qquad 2\pi \le \kappa_{total} \le \infty \qquad \text{(Eq.38)}$$

As the convex object will have the minimum value, a rough object will have a higher value.

4.16 Radial Distance Measures

Radial distance is the distance from the center of mass to the perimeter point (x_i, y_i) as shown in Fig. 4.16, redrawn from [Wirth, 2001]. So the radial distance is defined as:

$$d(i) = \sqrt{[x(i) - \bar{x}]^2 + [y(i) - \bar{y}]^2} \qquad i = 0,1,....N-1 \qquad \text{(Eq.39)}$$

Here $d(i)$ is a vector obtained by the distance measure of the boundary pixels. A normalised vector $r(i)$ is obtained by dividing $d(i)$ by the maximum value of $d(i)$. The vector $r(i)$ is used for calculating entropy and Fourier descriptor [Kilday et al, 1993].

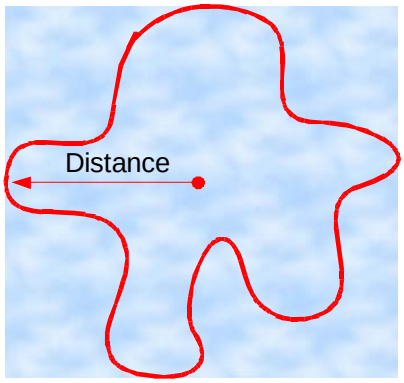


Figure 4.16 radial distance

4.16.1 Entropy

Entropy [Seul et al, 2000] as a measure of information in data is given by the sequence $r(n)$:

$$E = -\sum_{b=1}^{b} h_b \log h_b \qquad \text{(Eq.40)}$$

where h_b is the b-bin probability histogram which is the distribution of $r(n)$.

If we analyze the frequencies of the normalized radii by sorting them into a histogram with B bins, we get an indication whether the radii are mainly of the same size or whether they are oscillating between small and high values. If we have, for examples, 20 radii and we sort them into a histogram of size four, we get the probability of the sizes of length <0, 0.25) (bin1), <0.25, 0.5) (bin2), <0.5, 0.75) (bin3), <0.75, 1> (bin4). If the values of the bins are 5, 5, 5, 5 we have a mixture of all different sizes and a high entropy. If the shape is very regular, for example with all its radii between <0.5, 0.75) the bins are 0, 0, 20, 0 and we get a low entropy. We can measure the entropy by log to base 2. In this case, the first example will deliver: -0.25*(-2) - 0.25*(-2)-0.25*(-2)-0.25*(-2) = 2, a high entropy. In the second example, we have 0 + 0 -1*1 + 0 = -1, which is a low entropy.

4.16.2 Fourier Descriptor

Normalised radial distance $r(n)$ is analysed in spectral domain by discrete Fourier transform. The formula for DFT is:

$$a(u) = \frac{1}{N}\sum_{n=0}^{N-1} r(n) e^{-j2\pi mu/N} \quad u = 0,1,....., N-1 \qquad \text{(Eq.41)}$$

where $a(u)$ is the complex coefficient and is called the Fourier descriptor [Gonzalez and Woods, 2003]. The Fourier descriptor measures the regularity of a shape by analyzing its radial distances. The radial distances are ordered; say anticlockwise, and then the frequencies of these data are calculated. If, for example, higher Fourier coefficients of a(u) are close to zero, the shape radii have only low frequencies and are therefore quite regular. If the contrary is true, so the shape has a high fluctuation, i.e., it is quite irregular.

4.17 Linear Discriminant Analysis

LDA is one of the most popular statistical tools for classification. Although it is just a linear classifier, its use is motivated by its high stability and robustness. LDA searches for an optimal projection direction of the feature space into one dimension in such a way that the projected features of the different classes can be easily separated [Fukunaga, 1990]. In particular, in this case where the number of training samples is small compared to the dimension of the feature space it is much better than classifiers which work directly in the high dimensional feature space. Different features could have completely different value ranges. For the reason of numerical stability of the LDA, all features are scaled to the same value range [-1:1].

In the context of the project, LDA was mainly used for discriminating the classes. There are many other learning algorithms which will work in a similar way to construct a

decision plane or surface to classify cells. LDA constructs the decision plane as shown schematically in figure 4.17.

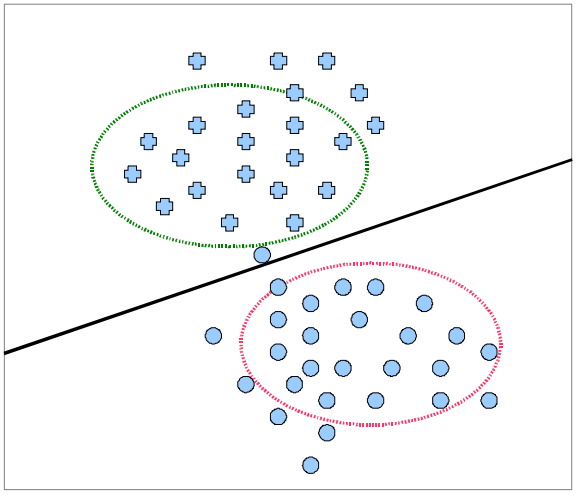


Figure 4.17 LDA separates two groups by a line

The features obtained from this thesis are independent of many learning algorithms so that other supervised and unsupervised pattern recognition techniques can be used. In this thesis an implementation of LDA was used which was developed at Fraunhofer-FIT institute.

5 Results and analysis

There are five different HeLa cell images selected for the experiment. The tasks are to calculate the feature values (described in Method section) of the cells and classify them according to their morphology. There is a large selection of morphological features which are used to classify the cells. As described earlier, it is possible to classify all types of cell-lines, but only HeLa cell images are presented here for the experiment. It should be observed that not all or the same types of features are required to classify different cells types. This is also true for the same type of cell line. For example, in the image in figure 5.1, six features were needed to correctly classify the cells. But for the correct classification of the cells in figure 5.9, eleven features were used. The following five experiments are based on five different images.

5.1 Experiment on the first image

Fig 5.1 shows the first image which is magnified 10 times its original size.

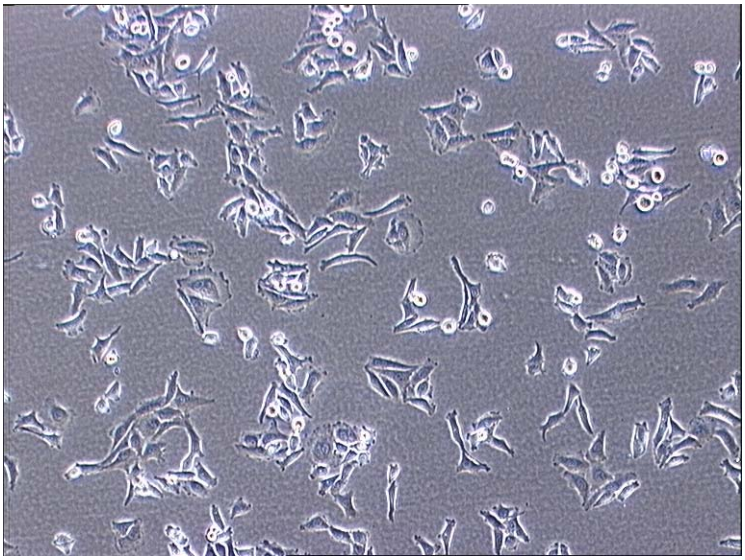


Figure 5.1 HeLa cells before segmentation

In Fig 5.1, there are two types of cells visible, one is round and the other has irregular elongated shape. The goal is to separate them automatically into two different classes and count their population.

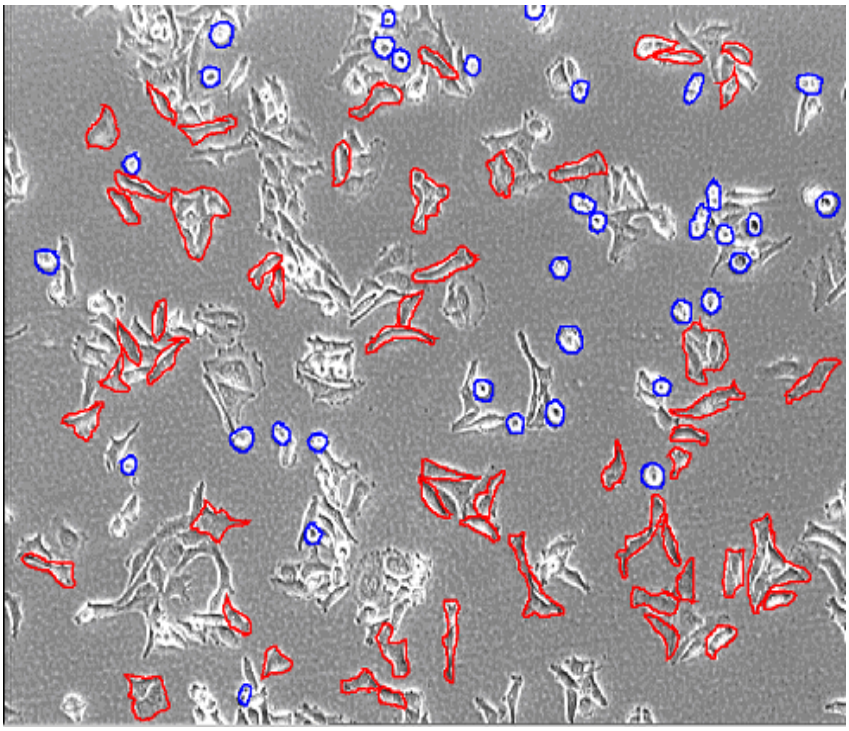


Figure 5.2 HeLa cells after segmentation, red color shows cell class A and a blue color shows cell class B

Cells in Fig 5.2 have been manually segmented to observe their properties. As described in the Introduction the scope of this thesis is feature selection and classification – therefore, manually selected cells are obtained for classification and thus for the experiment only a subset of cells are selected (manually) from a single image. The cells with irregular shape, colored red, are named **Cell Class A** and the spherically shaped cells, coloured blue, are named **Cell Class B**. Cells that belong to Cell Class A are attached to the substrate and the corresponding cells of Cell Class B are in the cleavage stage.

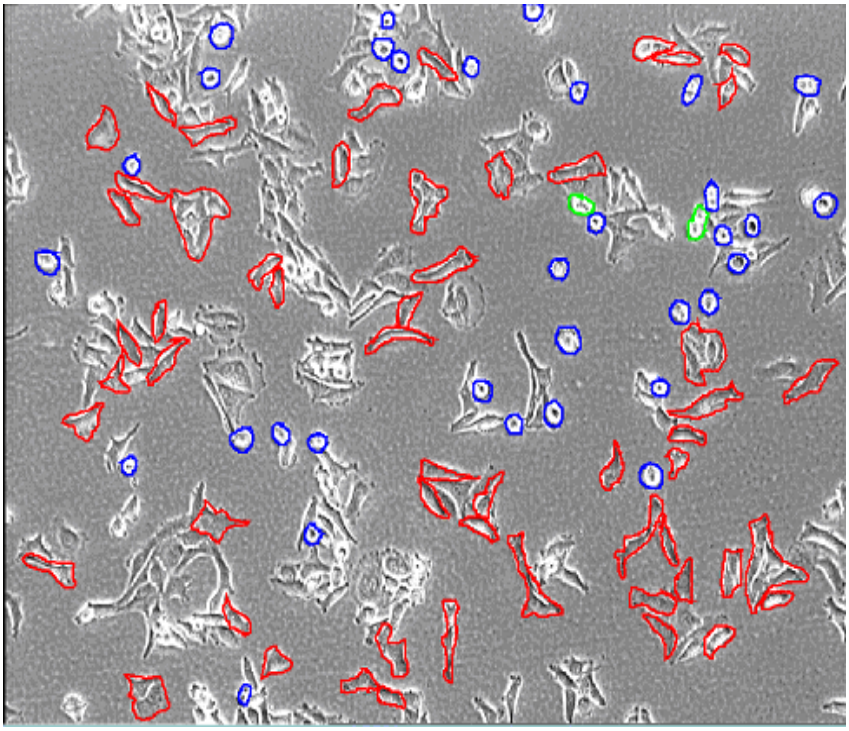


Figure 5.3 classified cells as Cell Class A (red), Cell Class B (blue) and misclassification in green

Fig 5.3 shows the result from the classification with LDA. The classification algorithm has chosen 6 features by leaving and adding shape features. These features yield the best classification result. They are EccentricityMomentFeature, VarianceFeature, SpreadMomentFeature, BendingEnergyFeature, AreaMomentFeature and CircularityFeature. In Cell Class A (red) there are 55 cells and in Cell Class B (blue) there are 35 cells.

The tables 5.1 and 5.2 give an overview of how the different features are distributed for the data of Cell Class A and Cell Class B in the form of mean and standard deviation. They are listed here to give a certain feeling about their discriminative power.

Different learning algorithms may make different use of these features. Some learning algorithms combine all features, but give them different weights depending on their discriminative quality. The stepwise discriminant analysis creates a subset of discriminating features. To achieve this, the scatter of the feature values around the group means (within class scatter) and scatter around the common mean (between class scatter) is calculated. Good features are those which maximize the between class scatter and minimize the within class scatter. In the algorithm the quotient of the between and within class scatter is calculated. At each iteration step, every feature is singly removed and the quotient is recalculated. If the change of the quotient is below a threshold, the feature stays removed. Earlier removed features are added to the feature-set and the quotient is recalculated. If the change of the quotient is above a threshold, the feature stays added. This is done until no features are either removed or added [Jennrich, 1977].

Table 5.1 shows the results for Cell Class A. Mean and standard deviation is calculated for each feature based on all 55 cells belonging to class A.

Cell Class A		
Feature	Mean	Standard deviation $(\pm)$
EccentricityMomentFeature	-0.8194	0.3725
CircularVarianceFeature	-0.2130	0.2619
SpreadMomentFeature	-0.7392	0.3695
BendingEnergyFeature	-0.5458	0.3571
AreaMomentFeature	-0.4233	0.3568
CircularityFeature	-0.3931	0.2939

Table 5.1 mean and standard deviation of feature values of Cell Class A of image 1

Table 5.2 shows mean and standard deviation for all involved features, based on the 35 cells which belong to Cell Class B.

Cell Class B		
Feature	Mean	Standard deviation $(\pm)$
EccentricityMomentFeature	-0.9993	0.0014
CircularVarianceFeature	-0.1252	0.3720
SpreadMomentFeature	-0.9706	0.0156
BendingEnergyFeature	-0.8070	0.0957
AreaMomentFeature	-0.7524	0.0682
CircularityFeature	0.6424	0.3068

Table 5.2 mean and standard deviation of feature values of Cell Class B of image 1

The mean and corresponding standard deviation shows that there is little overlapping between the values of the features. Other features are discarded because of overlapping values and only 6 features are enough to discriminate between Cell Class A and Cell Class B. This analysis shows the normal distribution given by

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{(Eq.45)}$$

Here, two randomly selected features named CircularityFeature and BendingEnergy Feature are considered for analysis. For the Circularity Feature the mean for Cell Class A is -0.3931 and the standard deviation is 0.2939, and the corresponding values for Cell Class B are 0.6424 and 0.3068. In Fig 5.4, the normal distribution curves show that these two groups are well separated as their means are far apart from each other and the data is well distributed around the mean for both of cell classes.

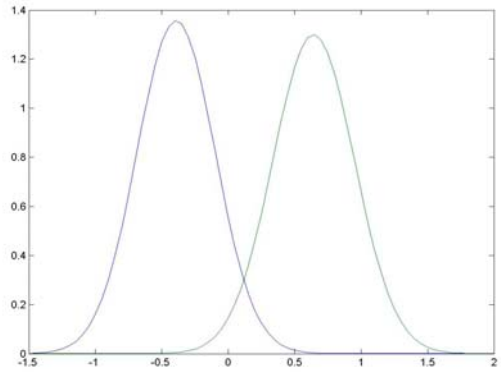


Figure 5.4 normal distribution of circularity feature of Cell Class A (left) and Cell Class B (right)

For the BendingEnergyFeature the mean and standard deviation for Cell Class A are -0.5458 and 0.3571 and for Cell Class B they are -0.8070 and 0.0957. These two groups are also well separated, although the means of two types of cell classes are somewhat closer than the circularity feature but still well separable (Fig.5.5).

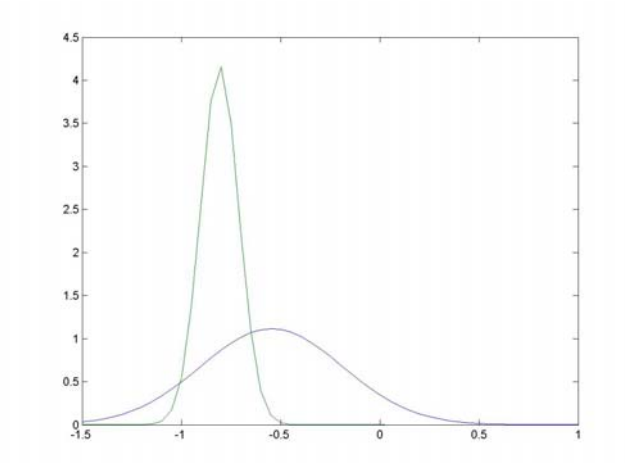


Figure 5.5 normal distribution of Bending Energy feature of Cell Class A (right) and Cell Class B (left)

Both the distribution figures 5.4 and 5.5 show that the Bending Energy Feature and Circularity feature are good classifiers for Cell Class A and Cell Class B though there are some small overlapping regions.

The classification result is shown in the Figure 5.3. There are two cells in Cell Class B that are wrongly classified as Cell Class A (green). The result is tabulated in Table 5.3 in the form of confusion matrix. The classification algorithm (LDA) randomly chooses thirty percent of the actual data for training. Thus 70% belongs to the test set. 38 cells of Cell Class A were used in the prediction, and all 38 cells were predicted correctly. This means that no cells of Cell Class A were predicted as type of being Cell Class B. Furthermore, of 24 cells of Cell Class B used in the prediction, 22 were classified correctly, while two were incorrectly predicted as Cell Class A (Table 5.3).

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	38	0
	Cell Class B	2	22

Table 5.3 prediction results for image 1

Table 5.4 shows the probability of each type of prediction. All cells belonging to Cell Class A are predicted 100% correctly, while the corresponding percentage for the cells belonging to Class B is almost 92%, meaning that 8% are classified incorrectly as shown in Table 5.4.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	1.0000	0.0000
	Cell Class B	0.0833	0.9167

Table 5.4 prediction statistics for image 1

5.2 Experiment on the second image

Fig 5.6 shows the second image from the experiment which is magnified 10 times its original size.

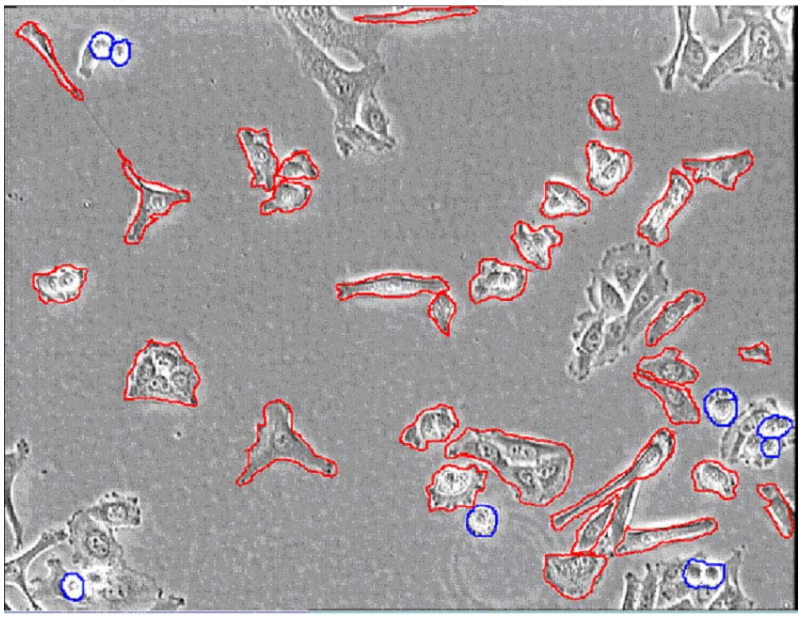


Figure 5.6 classified cells as Cell Class A (red), Cell Class B (blue) and no misclassification

For the image in figure 5.6 only the resulting classified image is shown as the original and segmented images give almost the same information.

There are 31 attached cells (Cell Class A) and their chosen features are tabulated in Table 5.5 along with mean and standard deviation.

Cell Class A		
Feature	Mean	Standard deviation $(\pm)$
CircularVarianceFeature	0.1938	0.3001
Eccentricity1MomentFeature	0.1417	0.4447
SolidityFeature	0.7250	0.2145
MinorAxisFeature	-0.1132	0.3902
CircularityFeature	-0.2156	0.3700

Table 5.5 mean and standard deviation of feature values of Cell Class A of image 2

There are only 8 round cells (Cell Class B) in Fig 5.6. The means and standard deviations of their corresponding features are shown in Table 5.6.

Cell Class B		
Feature	Mean	Standard deviation $(\pm)$
CircularVarianceFeature	0.3998	0.3586
Eccentricity1MomentFeature	0.6816	0.2477
SolidityFeature	0.9823	0.0183
MinorAxisFeature	-0.3423	0.1240
CircularityFeature	0.6630	0.2675

Table 5.6 mean and standard deviation of feature values of Cell Class B of image 2

Table 5.7 shows that there is no misclassification. 21 cells of Cell Class A were used in the classification and all were predicted correctly. All the 5 cells belonging to Class B was also correctly predicted.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	21	0
	Cell Class B	0	5

Table 5.7 prediction results for image 2

Table 5.8 shows the probability of each type of prediction. All cells belonging both to Cell Class A and to Cell Class B are 100% correctly classified. This means that there is no misclassification.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	1.0	0.0
	Cell Class B	0.0	1.0

Table 5.8 prediction statistics for image 2

5.3 Experiment on the third image

Fig 5.7 is the third image for the experiment which is magnified 20 times its original size.

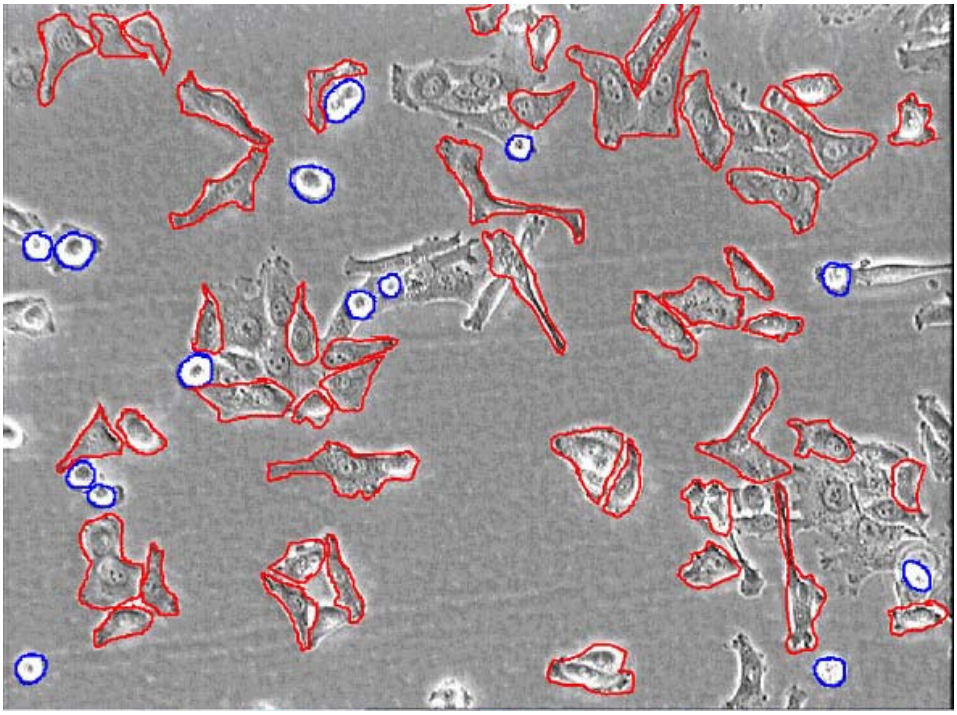


Figure 5.7 classified cells as Cell Class A (red), Cell Class B (blue) and no misclassification

For this experiment, the classification algorithm has chosen 11 features which yield the best classification result, by leaving and adding shape features. There are 47 irregular cells attached to the substrate which were segmented as red. Their means and standard deviations are given in Table 5.9.

Cell Class A		
Feature	Mean	Standard deviation $(\pm)$
ElongationFeature	-0.3061	0.3807
ConvexPerimeterFeature	-0.1182	0.3254
CircularVarianceFeature	0.0870	0.3963
RectangularityFeature	0.2390	0.3152
SpreadMomentFeature	-0.7362	0.3590
RoundnessFeature	0.4009	0.4250
BoundaryFollowFeature	0.7375	0.2992
RoundnessGravityFeature	0.0850	0.4225
MinorAxisFeature	-0.3102	0.2721
CircularityFeature	-0.3357	0.3117
MajorAxisFeature	-0.0924	0.3557
ConvexityFeature	0.8724	0.0745

Table 5.9 mean and standard deviation of feature values of Cell Class A of image 3

There are 14 round cells which are assumed to be in cleavage and floating and therefore belong to Cell Class B. The means and standard deviations of their features are given in Table 5.10.

Cell Class B		
Feature	Mean	Standard deviation $(\pm)$
ElongationFeature	-0.4032	0.0518
ConvexPerimeterFeature	-0.5991	0.0698
CircularVarianceFeature	0.2152	0.4164
RectangularityFeature	0.8749	0.0793
SpreadMomentFeature	-0.9793	0.0142
RoundnessFeature	0.9429	0.0380
BoundaryFollowFeature	0.6666	2.2204
RoundnessGravityFeature	0.8923	0.1117
MinorAxisFeature	-0.4979	0.0778
CircularityFeature	0.7792	0.1668
MajorAxisFeature	-0.6463	0.0724
ConvexityFeature	0.9452	0.0499

Table 5.10 mean and standard deviation of feature values of Cell Class B of image 3

Table 5.11 gives the result in the form of confusion matrix. There are 32 cells of Cell Class A which are also predicted in Cell Class A and there are 9 cells in Cell Class B also predicted in Cell Class B. Thus, the result contains no misclassification.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	32	0
	Cell Class B	0	9

Table 5.11 prediction results for image 3

For this image as well there is 100% correctly classified result was obtained for both the classes. This result is shown in Table 5.12.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	1.0	0.0
	Cell Class B	0.0	1.0

Table 5.12 prediction statistics for image 3

5.4 Experiment on the fourth image

Fig 5.8 shows the fourth image from the experiment which is magnified 20 times its original size.

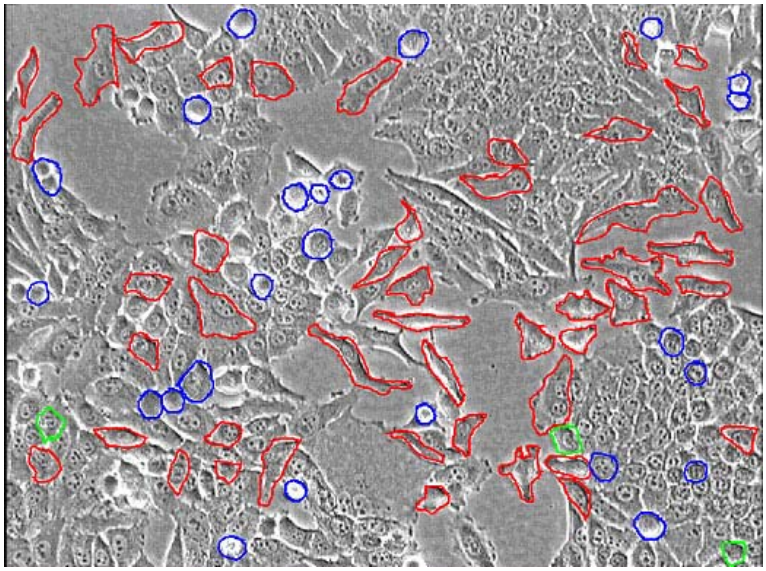


Figure 5.8 classified cells as Cell Class A (red), Cell Class B (blue) and misclassification in green

Linear Discriminant Analysis has chosen 7 features as best discriminating features for the fourth experiment. There are 49 cells of Cell Class A. Table 5.13 gives means and standard deviations of 7 features of Cell Class A.

Cell Class A		
Feature	Mean	Standard deviation (±)
TotalAbsoluteCurvatureFeature	0.7000	0.1921
CircularVarianceFeature	0.0588	0.3370
ConvexAreaFeature	-0.2863	0.4030
SpreadMomentFeature	-0.7463	0.3362
RoundnessGravityFeature	0.1031	0.4823
MinorAxisFeature	0.2692	0.3140
CircularityFeature	-0.2198	0.3567

Table 5.13 mean and standard deviation of feature values of Cell Class A of image 4

There are 24 cells of type Cell Class B. Means and standard deviations of 7 features of this type are given in the table 5.14.

Cell Class B		
Feature	Mean	Standard deviation (±)
TotalAbsoluteCurvatureFeature	0.6184	0.0629
CircularVarianceFeature	0.1059	0.2529
ConvexAreaFeature	-0.7020	0.1063
SpreadMomentFeature	-0.9617	0.0276
RoundnessGravityFeature	0.8554	0.0872
MinorAxisFeature	0.1441	0.1884
CircularityFeature	0.6795	0.1604

Table 5.14 mean & standard deviation of feature values of Cell Class B of image 4

Table 5.15 gives the classification result. There are 34 cells of type Cell Class A, among which 31 cells are classified in the same class but 3 cells are misclassified in the other class. There are 16 cells of type Cell Class B and they were predicted correctly.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	31	3
	Cell Class B	0	16

Table 5.15 prediction results for image 4

Table 5.16 shows the probability of prediction of cells in image 5.8. Cells belonging to Cell Class A are predicted 91% correctly, whereas the percentage of correct classification of cells belonging to Class B is 100%. 8% of Cell Class A are classified incorrectly. This result is shown in Table 5.16.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	0.9118	0.0882
	Cell Class B	0.0	1.0

Table 5.16 prediction statistics for image 4

5.5 Experiment on the fifth image

Fig 5.9 shows the fourth image from the experiment which is magnified 20 times its original size.

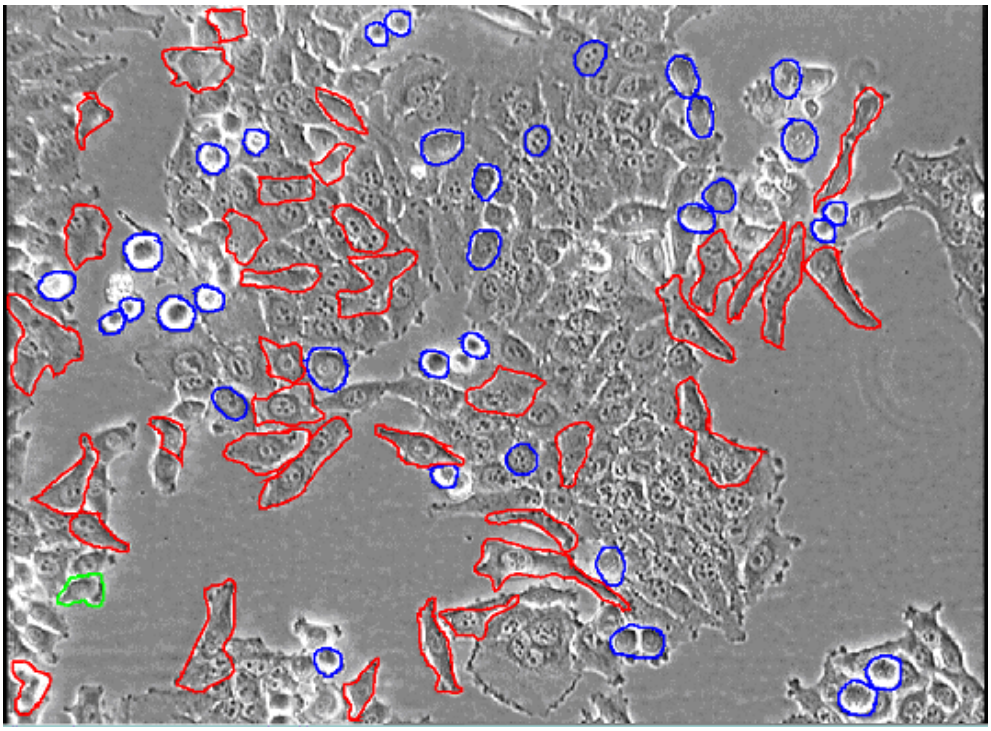


Figure 5.9 classified cells as Cell Class A (red), Cell Class B (blue) and misclassification in green

In the fifth and last experiment the classification algorithm found 11 features which are discriminating the two classes. There are 37 cells of irregular type of Cell Class A. Features with their corresponding means and standard deviations of this class are listed in Table 5.17.

Cell Class A		
Feature	Mean	Standard deviation (±)
TotalAbsoluteCurvatureFeature	0.6468	0.2149
EccentricityMomentFeature	-0.8661	0.3553
ConvexPerimeterFeature	0.0988	0.3119
CircularVarianceFeature	-0.2764	0.2151
RectangularityFeature	0.2310	0.3352
ConvexAreaFeature	-0.1967	0.4089
SolidityFeature	0.6822	0.1780
BoundaryFollowFeature	0.4980	0.1394
AreaMomentFeature	-0.0433	0.4151
CircularityFeature	-0.3572	0.3423
MajorAxisFeature	0.0310	0.3807

Table 5.17 mean and standard deviation of feature values of Cell Class A of image 5

There are 34 cells of Cell Class B. Features for those cells with means and standard deviations are given in the Table 5.18.

Cell Class B		
Feature	Mean	Standard deviation (±)
TotalAbsoluteCurvatureFeature	0.5979	0.0741
EccentricityMomentFeature	-0.9996	0.0012
ConvexPerimeterFeature	-0.4521	0.1048
CircularVarianceFeature	0.0041	0.4199
RectangularityFeature	0.8476	0.0852
ConvexAreaFeature	-0.7292	0.0950
SolidityFeature	0.9702	0.0186
BoundaryFollowFeature	0.4705	0.1011
AreaMomentFeature	-0.6104	0.1369
CircularityFeature	0.6386	0.2149
MajorAxisFeature	-0.5705	0.0911

Table 5.18 mean and standard deviation of feature values of Cell Class A of image 5

As described in the above experiment, there was 1 misclassification of Cell Class A in 25 cells and none in Cell Class B in 23 cells as shown in the Table 5.19.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	24	1
	Cell Class B	0	23

Table 5.19 prediction statistics for image 5

Among 25 five cells of Cell Class A, 24 are predicted in the same class that is 96% correct classification of Cell Class A and 4% are misclassified. In Cell Class B, 100% are correctly classified. These statistical results are given in the Table 5.20.

		Predicted	
		Cell Class A	Cell Class B
Actual	Cell Class A	0.9600	0.0400
	Cell Class B	0.0	1.0

Table 5.20 prediction statistics for image 5

6 Discussion and Conclusion

For a given cell line only a subset of cell features has real discriminative power. It can be noticed, for example; that only 5 to 12 features were used to classify the HeLa cells. There were 25 features in total for the experiments. Among those, 18 were actually used for classification. CircularityFeature and CircularVarianceFeature are used in all 5 experiments. 7 properties have no influence on this cell line. If lymphoblast or fibroblast cell types are under experiment, other morphological descriptors may be used from the database. This is because they will have different properties so they may need different descriptors.

Prediction results for the five HeLa images are very good:

- For the first image, 100% of cell Class A are predicted correctly, while 92% of Cell Class B are predicted correctly.
- For the second image, 100% of both types of cells were predicted in the right group, meaning that there is no misclassification at all.
- For the third image, there is again 100% correct classification.
- For the fourth image, 92% cells of Class A are predicted correctly. Cells in class B are predicted 100% correctly.
- For the last image, 96% of Class A are predicted correctly and the rate for Class B is 100%.

The result is better when one of the cell types is in minority as seen in the second and third image. More or less, this will always be the case as both types cannot be the same or even nearly equivalent in number as cells in cleavage are fewer than cells in non-cleavage.

The misclassification rate may be reduced in a successive run of the classification algorithm as experienced during the experiments by the writer of this paper. The algorithm randomly chooses 30% of the training data. It may choose better data or the worse ones in different successive runs but the result does not vary drastically. Only 1 to 3% of variation occurs, which is tolerable. Suppose if five successive runs of the classification algorithm are taken, the best run could be saved for the best result.

On average, the prediction accuracy is 98% which is at least 3 percentage points better than previously reported in the literature. As described in the paper ‘Classification of cultured mammalian cells by shape analysis and pattern recognition’ [Olson et al., 1980], when LDA or other machine learning techniques were used, as the feature dimensionality increased, the performance of the classifier either decreased or remained the same. But the LDA used for this experiment uses the property of leaving and adding those features which yield the best classification result. Exact mathematical definitions are always necessary for repeatability of the experiment. Any black box features with unknown definition were not used. Also there are some new features added as cited in the end of section 2.4 Related Work.

In the context of classification, cross validation can be used to check the goodness of a classification by removing one or more training data from the training set and performing LDA with the reduced training set and checking then, how well LDA will perform on the leaving-out data. Here we have not done this, but we have instead evaluated (classified) the whole cell image by hand and used this hand-made classification as a template to evaluate the result delivered by LDA and got a good performance of 98%.

Among new features which we used in this work, the circularity and circular variance features were used in all five experiments. As described above, they are the most common and widely used features which were most important for classification of the two kinds of cells, that is Cell Class A and Cell Class B. As these two kinds of cells are mainly classified based on whether they are round or not, it is quite intutive that these two features should be present in all five experiments. EccentricityMomentFeature and SpreadMomentFeature were used in three experiments. The experiment with image1, image3 and image 4 used SpreadMomentFeature and the experiments with image1, image2 and image 5 used the EccentricityMomentFeature. The RoundnessFeature was used in the experiments with image3 and image4. The TotalAbsoluteCurvatureFeature was used in the experiments with image 4 and image 5. Other important features which contributed to cell classification were the ConvexAreaFeature, the ConvexPerimeterFeature, the ConvexityFeature and the BendingEnergyFeature. Unused features may be used in other runs of the program. Some features may not be useful for one type of cell but will be useful for a different cell line.

HeLa cells are elongated and stay attached to the growing plate, except a few rounded cells in cleavage. Other cell types may have completely different characteristics and these shape-descriptors can be adapted to that specific cell type. One important issue is to set the proportion of the cells in cleavage and cells on normal growth. But when a good discriminating result is gained, there is no problem to set a threshold value accordingly (depending on cell type) by the user.

There is some computational overhead with feature classification. There are some descriptors which have no discriminatory effect on classification for specific cell lines but will be there in the repository. Each time when LDA classifies the cells those descriptors will also be considered for evaluation which is not necessary. To make the system capable to perform classification on all three kinds of cells, it is necessary to add some more features to the feature set. This will again cost an extra overhead.

7 Future work

Automatic image segmentation

Cell images are here segmented manually. But in real time systems it has to be automatic and also thousands of cell segmentations have to be done in seconds. This is a challenging task because images are noisy. There needs to be lot of pre-processing even before segmentation. A method which is successful in many areas of image processing is called the watershed transformation algorithm [Pitas, 2000]. In particular, a very powerful gray-scale segmentation methodology results from applying the watershed to the morphological gradient of an image to be segmented. The watershed algorithm splits the image into regions similar to the drainage regions of a landscape. If the intensity of the image is interpreted as elevation in a landscape, the watershed algorithm can be used to find mountains, lakes and catchment basins in the landscape. This can serve as a ground for segmentation algorithms.

More cell descriptors can be added in case of future need

This experiment was done on HeLa cell images which is a epithelial-like cell line. The features used for this experiment work fine to classify this kind of cells. It is also hoped that the features used here will also work for other cell lines because very general descriptors were used. However, this cannot be confirmed until other kinds of cell lines are actually used for the experiment. It is anticipated that in the future other cell lines will be used to test the classification result, and if needed other features will be added on demand.

Overhead reduction

As described in the previous section, Discussion and conclusion that there will be computational overhead if more and more features are added. Some features will classify Epithelial-like cells better and some will classify Lymphoblast-like cells better. It is not necessary to calculate all of the features in the database. The classification algorithm can be trained for different cell lines so that every cell line will have a specific list of relevant features. In this way much of the computational overhead can be avoided. This can be incorporated in the future when other cells are used in the experiment.

8 References

[Achard et al, 2000] Achard, C., Devars, J., Lacassagne, L., 2000. Object Image Retrieval with Image Compactness Vectors. *Proceedings of the International Conference on Pattern Recognition (ICPR'00)* 1051-4651/00

[Agouris and Stefanidis, 2000] Agouris, P., Stefanidis, A., 2000. *Integrated Spatial Databases Digital Images and GIS.* Portland, ME:Springer.

[Alberts et al., 1998] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walte,r P. 1998 *Essential Cell Biology,* New York:Garland

[Ballard and Brawn, 1982] Ballard, D. and Brown, C. 1982. *Computer Vision.* Englewood Clifffs, NJ: Prentice Hall

[Cann, 2000] Cann, A. J., 2000. *Virus Culture*. Oxford:Oxford University Press.

[Castleman, 1996] Castleman, K. R., 1996. *Digital Image Processing.* Upper Saddle River, NJ:Prentice Hall.

[Celebi and Aslandogan, 2005] Celebi, M. E., Aslandogan, Y. A., 2005. A Comparative Study of Three Moment-Based Shape Descriptors. *International Conference on Information Technology: Coding and Computing (ITCC'05).* 1, 788-793.

[Costa and Cesar, 2000] Costa, L. F., Cesar, R. M., 2000. *Shape Analysis and Classification.* Boca Raton, Florida:CRC Press.

[Dawe et al., 1994] Dawe, R. K, Sedat, J. W, Agard, D. A and Cande, W. Z. 1994. Meitotic chromosome pairing in maize is associated with a novel chromatin organisation. *Cell* 76:901-902.

[Derry, 2002] Derry, G. N., 2002. *What Science Is and How It Works*. Princeton, NJ:Princeton University Press.

[Duda et al., 2001] Duda, R. O., Hart, P.E., Stork, D. G., 2001. *Pattern Classification*. 2nd ed. New York, NY:Wiley Interscience.

[Fukunaga, 1990] Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition.* 2nd ed. San Diego, CA:Academic Press.

[Gonzalez and Wintz, 1987] Gonzalez, R.C., Wintz, P., 1987. *Digital Image Process*ing. Reading, MA:Addisson-Wesley.

[Gonzalez and Woods, 2003] Gonzalez, R. C., Woods, R. E., 2003. *Digital Image Processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall.

[Hann and Oprea, 2004] Hann, M.M., Oprea, T.I. 2004. Pursuing the leadlikeness concept in parmaceutical research. *Current Opinion in Chemical Biology.* 8:255-263.

[Hu, 1962] Hu, M. K., 1962 Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*. 17-8 (2), 197-187.

[Jain, 1989] Jain, A. K., 1989 *Funadmental of Digital Image Processing.* Englewood Cliffs. NJ:Prentice-Hall.

[Jenkin, 1997] Jenkin, M., 1997. *Computational and Psychophysical Mechanisms of Visual Coding.* Cambridge:CambridgeUniversity Press.

[Jennrich, 1977] Jennrich, R.I., 1977. Stepwise discriminant analysis. *Statistical methods for digital computers (K. Enslein, A. Ralston and H.S. Wilf, Eds)* 76-95. New York, NY: John Wiley.

[Kilday et at., 1993] Kilday, J., Palmieri, F., and Fox, M.D., 1993. Classifying mammographic lesions using computerized image analysis. *IEEE Transactions on Medical Imaging.* 12, 664-669.

[Kim et al., 2002] Kim, S., Kim, J., Kim, S., Kim, M., 2002. Usefulness of Boundary Sequences in Computing Shape Features for Arbitrary Shaped Regions. *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02)* 1051-4651

[Lee et al., 2003] Lee, S. W., Bulthoff, H. H., Poggio, T., 2003. *Biologically Motivated Computer Vision*. Seoul:Springer.

[Levine, 1985] Levine, M. D., 1985. *Vision in man and machine.* New York, NY:McGraw-Hill.

[Lindblad, 2003] Lindblad, J., 2003. *Development of Algorithms for Digital Image Cytometry*. Thesis (PhD). Uppsala University.

[Manning and Schütze, 1999] Manning, C. D., and Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. Massachusetts, MA: MIT Press.

[Naik, 1998] Naik, R., 1998. *Creating classification features for biological images.* Thesis [Master's]. University of Minnesota Duluth.

[Olson et al., 1980] Olson, A. C., Larson, N. M., and Heckman, C. A., 1980. Classification of cultured mammalian cells by shape analysis and pattern recognition. *Proc. Natl. Acad. Sci. USA.* 77 (3) 1516-1520.

[Pitas, 2000] Pitas, I., 2003. *Digital Image Processing Algorithms and Applications*. New York, NY: John-Wiley & Sons.

[Rodenacker and Bengtsson, 2003] Rodenacker, K., Bengtsson, E., 2003. A feature set for cytometry on digitized microscopic images. *Analytical Cellular Pathology*. 25:1-36

[Ryan, 2003] Ryan, J. A. 2003. *Introduction to Animal Cell Culture* [online]. Acton, Corning Incorporated. Available from: http://www.corning.com/lifesciences/technical_information/techdocs/intro_animal_cell_culture.pdf [Accessed 17 August, 2005].

[Seul et al, 2000] Seul, M., Sammon, M. J., O'Gorman, L., 2000. *Practical Algorithms for Image Analysis*. Canbridge:Cambridge University Press.

[Shipley and Kellman, 2001] Shipley, T. F., Kellman, P. J., 2001. *From Fragments to Objects*. Amsterdam:Elsevier

[Turner et al., 1993] Turner, M., Austin, J., Allinson, N., and Thompson, P. 1993. Chromosome location and feature extraction using neural networks. *Image and Vision Computing* 11:235-239.

[Wied et al., 1989] Wied, G. L., Bartels, P. H., Bibbo, M. and Dytch, H. E. 1989. Image analysis in quantitative cytopathology and histopathology. *Human Pathology* 20:549-571.

[Wirth, 2001] Wirth, M. A., 2001. *Shape Analysis and Measurement* [online]. University of Guelph. Available from: http://hebb.cis.uoguelph.ca/~mwirth/Teaching/CIS6320/Lecture10.pdf [Accessed 13, March, 2005]

[Wittekind and Schulte 1987] Wittekind, C., and Schulte, E. 1987. Computerized morphometic image analysis of cytologic nuclear parameters in breast cancer. *Anal. Quant.* Cytol. And Hist. 9:480-484.

[Wohlberg et al., 1993] Wohlberg, W. H., Street, W. N., Mangasarian, O. L. 1993. Breast cytology diagnosis via digital image analysis. *Analy. Quant. Cytol. And Hist*. 15:396-404.

[Wohlberg et al., 1995] Wohlberg, W. H., Street, W. N., Mangasarian O. L. 1995. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analy. Quant. Cytol. And Hist*. 17:77-87.

[Wong and Hall, 1978] Wong, R. Y., Hall, E. L., 1978. Scene matching with invariant moments. *Computer Graphics and Image Processing*. 8 (1), 16-24.

[Ya, 2003] Ya, W. C., 2003. *Handbook of Fluidization and Fluid-Particle Systems*. New York, NY:Marcel Dekker.