



Primary Research

The detection of cervical cancer disease using an adaptive thresholding method through digital image processing

Eggi I. Putri ^{1,*}, Rita Magdalena ^{2,*}, Ledy Novamizanti ^{3,*}

^{1, 2, 3} Telkom Engineering School (TES)–Telkom University (Tel-U), Bandung, Jawa Barat 40257, Indonesia

Keywords:

Cervical cancer Pap smear
Digital image processing
Thresholding

Received: 15 January 2015

Accepted: 26 July 2015

Published: 15 October 2015

Abstract—Cervical cancer is a kind of cancer disease caused by human papillomavirus types 16 and 18 attacking women's cervix. To detect cervical cancer, the frequently-used method is Pap-Smear; however, errors often occur when the method is taken to diagnose the level of cervical cancer. Thus, a proper system is required, which is supposed to help identify the result of Pap-Smear. This study aims at designing a system to detect the symptoms of cervical cancer using MATLAB to solve these errors. The image processing begins with converting the type of image, followed by thresholding and noise removal using filters until the image has become ready to be detected. For a thresholding process, an Adaptive Thresholding method is taken, in which the thresholding focuses on local threshold values. The system can classify images into two types, i.e., normal and abnormal (precancerous). Abnormal type is divided into three subtypes, i.e., mild, moderate, and severe. An experiment is conducted on the proposed system, in which it is supposed to analyze 500 test images, including 250 for training and 250 for testing. A perfect 100% accuracy rate is obtained based on the testing process, while the average processing time is 25.4 seconds with a WS value at ten and a C value at -2.

© 2015 The Author(s). Published by TAF Publishing.

INTRODUCTION

The reproductive system in woman body consists of several parts, in which one of them is uterus. Uterus consists of cervix. The cervix connects the uterus with female reproductive system. In fact, cervical cancer is kind of cancer diseases caused by human papilloma virus type 16 and 18 attacking woman cervix. Cervical cancer is a frightening specter for every woman, because its

*Corresponding author: E.J. Putri, R. Magdalena and L. Novamizanti
E-mail: eggiintanputri@gmail.com; ritamagdalen@telkomuniversity.ac.id; ledyaldn@telkomuniversity.ac.id



symptoms is not clearly observable before an advanced stage. Thus, an early detection is regularly required to preventing and treating cervical cancer symptoms. The current cervical cancer detection method that is frequently used is Pap-Smear. However, errors often occur in diagnosing the level of cervical cancer. Therefore, a system that is able to help identifying the result of Pap-Smear is required. This research aims at providing a digital alternative in the early examination of cervical cancer by using software to help medical examiners to interpret Pap-smear results. Prior researches have attempted to use an evolving ANN technique [1], highly efficient second-order neural network training algorithms [2], and a neuro-fuzzy classification that is able to classify two classes, i.e. normal and abnormal, with a 75% average accuracy. On the other hand, the current research uses Adaptive Thresholding and K-Nearest Neighbor for classifying the cell types of cervical cancer. The classification is divided into two classes, i.e. normal and abnormal. The normal class is divided into four classes, i.e. columnar, parabasal, intermediate, and superficial. The abnormal class is divided into three classes, e.g. mild, moderate, and severe.

METHODOLOGY

Cervix

The reproductive system in a woman body consists of uterus, cervix, fallopian tube, and ovary. Uterus is located between hips and fallopian tube, while ovary is located at right and left positions (Figure 1) [3]. Cervix comes from a Latin term, referring to an entrance to uterus or the lowest part of uterus that patches to vagina [4].



FIGURE 1. Part of reproductive system in woman body [4]

Cervical Cancer

This type of cancer is known as a malignant tumor that attacks cervix. In a general situation, cervical cancer attacks women aged between 35-55 years old. In fact, almost 90% of cervical cancer begin from squamous cells in cervix, while the rests occur from gland cells inside the cervix line to uterus. Factors for a cervical cancer include human papilloma virus (HPV), smoking,

sexual intercourse at a young age, repeated changes of sexual partners, and herpes genitalis infection [4].

Types of Cervix Cells

Cells changing in cervix occurs slowly and it may not produce cancer cells. In fact, there is a pre-cancer/cervix dysplasia condition, in which there is a change from normal cells to abnormal ones before transform into a cervical cancer. In general, dysplasia cells have a big size, dark colored and stick with other cells in cervix.

The existence of pre-cancer cells must be carefully treated during the pre-cancer condition because a good treatment may possibly prevent pre-cancer cells to be cancer cells. There are seven types of cells in the pap-smear results taken in the current research (Figure 2). Besides, there are two types of cells, i.e. normal cells and abnormal ones (dysplasia) [1, 2]:

1. Normal cells

a. Columnar epithelial cells

This cell has nucleus and cytoplasm. The nucleus is located under cytoplasm. Looking from the top, nucleus is wider than cytoplasm area. From the side, cytoplasm looks wider. The width of nucleus area is $\sim 50 \mu\text{m}^2$, while it is darker than around cytoplasm.

b. Parabasal squamous epithelial cells

Nucleus in this type is shaped like a circle with $\sim 50 \mu\text{m}^2$ width, while the cytoplasm has $200\text{-}300 \mu\text{m}^2$ width.

c. Intermediate squamous epithelial cells

Nucleus in this cell has $20\text{-}35 \mu\text{m}^2$ width, which is narrower than in parabasal type. However, it has a larger cytoplasm with $800\text{-}1600 \mu\text{m}^2$ width.

d. Superficial squamous epithelial cells

Nucleus in this cell type has the same width with intermediate type ($20\text{-}35 \mu\text{m}^2$) and the same width of cytoplasm ($800\text{-}1600 \mu\text{m}^2$).

2. Abnormal cells (dysplasia)

a. Mild non-keratinizing dysplastic cells

This type is darker and has a larger width of nucleus than normal cells [1].

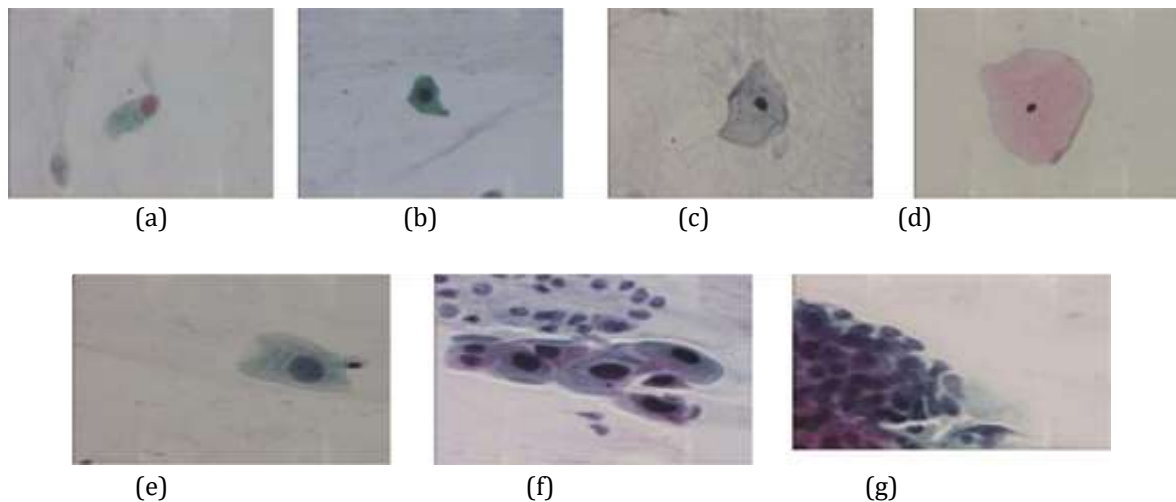


FIGURE 2. Cells in cervix: **(a)** Columnar epithelial cells; **(b)** Parabasal squamous epithelial cells; **(c)** intermediate squamous epithelial cells; **(d)** Superficial squamous epithelial cells; **(e)** Mild non-keratinizing dysplastic cells; **(f)** Moderate non-keratinizing dysplastic cells; **(g)** Severe non-keratinizing dysplastic cells

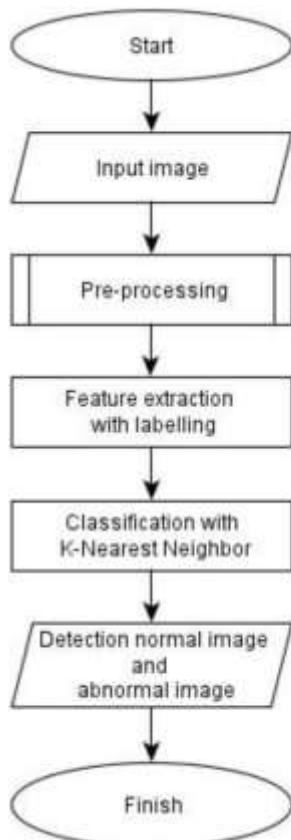


FIGURE 4. System process

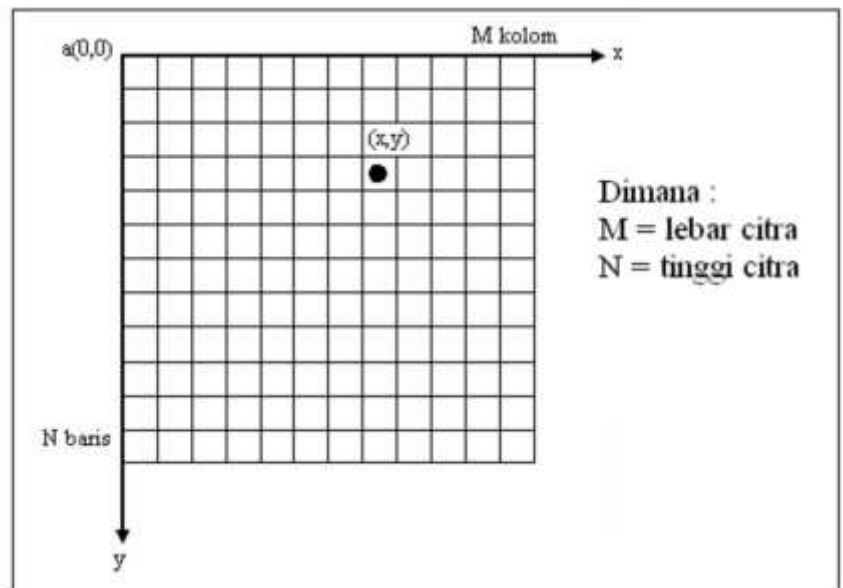


FIGURE 3. Representation of a digital image in computer

b. Moderate non-keratinizing dysplastic cells

This type has a darker color and a wider as well as larger nucleus. In this type, nucleus may change its shape to become badly. There are small circles in the nucleus [1].

c. Severe non-keratinizing dysplastic cells

This is the last conditional change cells for the pre-cancer. Its nucleus is dark, big, badly-shaped, in which nucleuses may form a group. In this type, cytoplasm has a dark color and a smaller size than nucleus [1].

Digital Image Processing

In general, digital image processing is the processing of a two-dimension image by using computer software [5]. According to Efford [6], image processing offers various techniques for manipulating and modifying an image [7-9]. The first is by converting image-to-image, for example, from digital camera, digital microscope, or scanner to a computer. An image with two-dimension data is represented by a $M \times N$ matrix, where M is column and N refers to row. The smallest element in an image is pixel. Pixel is a sample that has an intensity represented by an integer. A digital image is a pixel

collection stacked into a two-line dimension, in which row and column index (x,y) from a pixel is interpreted as an integer. Pixel (0,0) is located at top-left corner, while x-index moves to right and y-index moves to lower position. Figure 3 exhibits a representation of digital image in a computer.

Adaptive Thresholding Method

The thresholding process produces binary images, which have two grayscale values, *i.e.* black and white. Meanwhile, the adaptive thresholding is a threshold process that utilizes a local threshold, which is calculated based on neighboring pixel statistics. At least, there are three types of adaptive thresholding [7], in which the current research utilizes the first type (mean concerning local intensity).

1. Mean concerning local intensity

$$T = \frac{\sum_{(y,x) \in W} f(i,j)}{N_w} - C \quad (E.1)$$

where:

W = size for window in image

Nw = pixel in window

C = constant

2. Median approximation

$$T = \text{median} ((y, x), (y, x) \in W) - C \quad (E.2)$$

3. Maximum and minimum statistical approximation

$$T = \frac{\max(f(y,x), (y,x) \in W) + \min(f(y,x), (y,x) \in W)}{2} \quad (E.3)$$

RESULTS

System Design

The input is a cell image of cervical cancer from Pap-Smear results, in which all are in color and BMP format (goo.gl/zJms00). This research uses 500 images, in which 250 are training images and 250 are testing images. Figure 4 describes the process.

Pre-Processing

RGB to Grayscale. The input images are sampled into RGB format. The first process includes changing image format from RGB to grayscale. It is intended to enhance the density of information for next processes. The images are transformed from RGB (3

$$\text{Grayscale} = (0.299 + 0.587 + 0.114) \quad (E.4)$$

layers) to grayscale (1 layer) based on the following formula:

Adaptive Thresholding. This research uses adaptive thresholding, because there is a lighting difference on certain area between images. The method suggests a threshold value from sub-images, which acts as the local threshold. The threshold value is valid for sub-images in which the area in an image is given by window size value (WS) and constant (C). To achieve a better accuracy, WS value and C value are adjusted over certain areas.

Filtering. The current research uses a filter (bwareaopen) to remove noises within an image. In the proposed system, noises are removed with a restriction of wide white pixel value in image at less than 100. The value is set

by considering that a value less than 100 may not produce the best result, in which there are noises such as small circles. However, a value bigger than 100 may make the edge of an image to fade and break, and any object in the image cannot be correctly determined.

Labelling. Labelling is taken to check white pixels in an image. The checking is taken from left to right and upper to lower. It produces easy detection and classification. The labelling process chooses 30 larger wide cells in an image. To do the detection process, this research uses K-Nearest Neighbor

$$L_2(X, Y) = \sqrt{\sum_{i=1}^d (X_i - Y_i)^2}$$

(KNN) with a Euclidean Distance rule through the following formula:

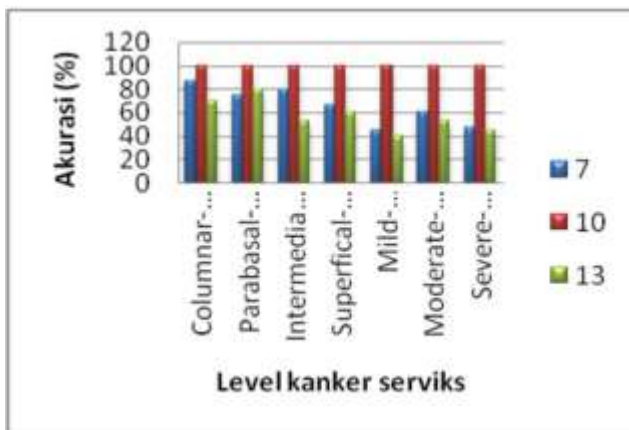


FIGURE 5. Testing for WS value in training images

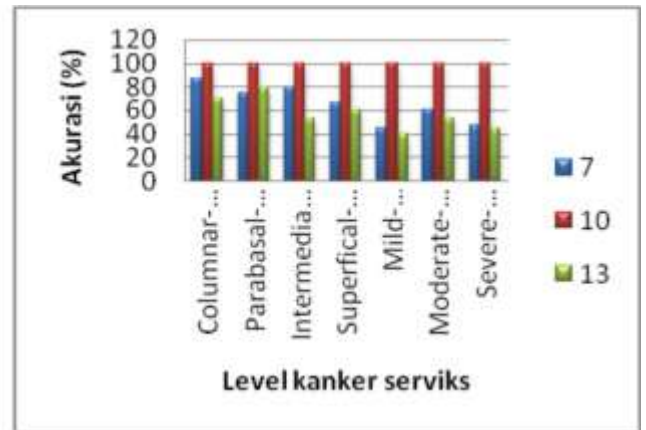


FIGURE 6. Testing for C value in training images

DISCUSSION

The input is an image of cervical cancer cell (goo.gl/zJms00), in which it is the result of a Pap-Smear test with a 768x568 pixels resolution in BMP format. There are 250 training images and 250 testing images for training and testing the system in non-real time. To determine the performance of the proposed system, it is tested through several scenarios:

- Varying the value of window size in adaptive thresholding
- Varying the value of constant (c) in adaptive thresholding
- Testing the system accuracy
- Testing the system computation time

Varying the value of window size in adaptive thresholding. During the testing, the value of window size in adaptive thresholding method is adjusted. The values are set at 7, 10 and 13 for every training image being tested (Figure 5). In the end, the best value for window size is found at 10.

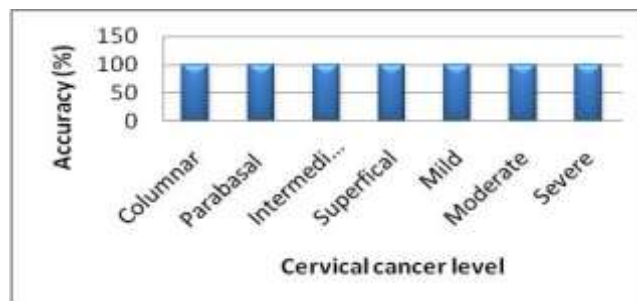


FIGURE 7. The result testing system from testing images Varying the value of Constant (C) in adaptive thresholding. In testing the system, the constant (C) value is adjusted in adaptive thresholding method. The values are set at -4, -2 and -6 for every training image (Figure 6). In the end, the best C value is at -2 for 250 training samples. Testing the system accuracy. Testing process is conducted with adjusted WS (10) and C (-2) over 250 testing images. Figure 7 shows the results, in which the system is able to classify with a perfect accuracy for all levels of cervical cancer. Testing the system computation time. From the testing over training and testing images in adaptive thresholding method, the longest time for a detection is 36.063 seconds. Over 250 testing images, the average time for a detection is at 25.4 seconds to classify a cervical cancer.

CONCLUSION

The designed system is able to detect cervical cancer by classifying 7 classes of the disease. The normal class is divided into four classes, while the abnormal class is divided into three classes with a good accuracy. The best value for window size (ws) is 10 and constant (C) is -2, while the wide of white pixels is 100 in adaptive thresholding method. The input is an image of cervix cell (goo.gl/zJms00), in which the image is taken from the result of a cervical cancer check-up by using Pap-Smear method with a 768x568 pixels resolution in BMP format. Then, the accuracy of the proposed system is perfect (100%) and it has a 25.4 seconds average time of detection with adaptive thresholding method.

REFERENCES

1. Kahfi RA. *Mendeteksi penyakit kanker serviks dengan menggunakan teknik evolving ANN*. [BSc thesis]. Bandung, ID: Telkom University; 2013.
2. Ampazis N, Perantonis S. Two highly efficient second-order algorithms for training feedforward networks. *IEEE Trans Neural Netw.* 2002; 13(5): 1064-1074. DOI: [10.1109/TNN.2002.1031939](https://doi.org/10.1109/TNN.2002.1031939)
3. McCormick CC, Giuntoli II RL. *John Hopkins medicine: Patients' guide to cervical cancer*. Ontario, CA: Jones & Bartlett Publishers; 2010.
4. Prayitno S. *Buku lengkap kesehatan organ reproduksi wanita*. Yogyakarta, ID: Saufa; 2014.
5. Jain AK. *Fundamentals of digital image processing*. Upper Saddle River, US: Prentice-Hall; 1989.
6. Efford N. *Digital image processing: A practical introduction using Java*. Boston, US: Addison-Wesley Longman Publishing; 2000.
7. Kadir A, Susanto A. *Teori dan aplikasi pengolahan citra*. Yogyakarta, ID: ANDI Publishers; 2013.
8. Putra D. *Pengolahan citra digital*. Yogyakarta, ID: Andi Publishers; 2010.
9. Nugrahanto. *Deteksi gejala hematuria berdasarkan jumlah eritrosit pada urine berbasis pengolahan citra* [BSc thesis]. Bandung, ID: Telkom University; 2012.
10. Jantzen J, George D. *Analysis of Pap-smear image data*. Paper presented at: Nature-Inspired Smart Information Systems (NISIS) 2nd Annual Symposium; 2006 November 29 - December 1; Tenerife, ES.
11. Sianipar RH. *Pemrograman MATLAB dalam contoh dan penerapan*. Bandung, ID: Informatika; 2013.

— This article does not have any appendix. —