BIOLOGICAL VIRUSES IMAGES RECOGNITION USING ARTIFICIAL INTELLIGENCE CLASSIFIER

AFIQ BIN AHMAD SHAKRI

UNIVERSITI MALAYSIA PERLIS
2018

BIOLOGICAL VIRUSES IMAGES RECOGNITION USING ARTIFICIAL INTELLIGENCE CLASSIFIER

by

AFIQ BIN AHMAD SHAKRI
1630212139

A thesis submitted in fulfillment of the requirements for the degree of Master of Science (Computer Engineering)

Faculty of Engineering Technology
UNIVERSITI MALAYSIA PERLIS

2018

ACKNOWLEDGEMENTS

First of all, thank you Allah Subhanahuwa Ta'ala for allowing me to made it to this point of my life. His blessings and guidance bestowed upon me has helped carry me all the way to this stage of my life.

I would like to extend my eternal thanks to my supervisor Dr. Muhammad Naufal bin Mansor of the Faculty of Engineering Technology at Universiti Malaysia Perlis. The door to Dr. Muhammad Naufal office and home was always open whenever I ran into a problem or had some question about my research and thesis writing during my research work period. Dr. Muhammad Naufal consistently allowed this paper to be my own work and at the same time steered me in the right direction whenever he thought I needed it. He was always there to help whenever I stumble during completing my research. Without his help I would not even reach the final stage of my research.

A very special gratitude goes out to my co-supervisor Dr. Ahmad Kadri bin Junoh for his help and also to all of the lecturers of PTT-106 Microbiology and PTT-207 Biomolecular and Genetic Engineering of academic session 2016/2017 for being patience with me and helping me completed the additional requirement classes for my research programme.

I would also like to thanks the Unimap Centre for Graduate Studies Staff for their help throughout the course of my research.

Finally, I must express my very profound gratitude to my parents and siblings for putting up with me and providing me with eternal support and continuous encouragement throughout my tears of study and during the process of researching and writing this thesis. I daresay this accomplishment would not have been possible without them and all of the person I mentioned here. Thank you.

TABLE OF CONTENT

PAGE

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIRS	Artificial Immune Recognition System
AIBN	Artificial Immune B-Cell Network
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
DNA	Deoxyribonucleic acid
DT	Decision Tree
FFT	Fast Fourier Transformation
F-kNN	Fuzzy k-Nearest Neighbor
FP-ANN	Feed Forward Back Propagation Artificial Neural Network
GE	Global Frequency Estimation of Prior Probability
GLCM	Gray Level Co-Occurrence Matrix
ISODATA	Iterative Self-Organizing Data Analysis Technique
KNN	K Nearest-Neighbor
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LE	Local Frequency of the Prior Probability
LLC	Locality-Constrained Linear Coding
MLP	Multi Layer Perceptron
MRI	Magnetic Resonance Imaging
NN	Neural Network
PCA	Principal Component Analysis
PCANet	Principal Component Analysis Network
PNN	Probabilistic Neural Network
RNA	Ribonucleic acid

SOM	Self-Organizing Map
SVM	Support Vector Machines
TEM	Transmission Electron Microscopy
UAIC	Unsupervised Artificial Immune Classifier
XOR	The Exclusive-Or

LIST OF SYMBOLS

Mk	Weighting factors
$d_E(x, y)$	Euclidean Distance on (x,y) coordinate
υ_i	Number of pixels
Ψ	Small neighborhood
Ω	Image domain
λ	Smoothness constraint
m_L	Local Mean
Φ_i	Vector Image
ω_k	Eigen-Vector
δ_k	Output Neuron
δ_j	Hidden Neuron
μ^2	Uniform Patterns
v_{ij}	Weight between Input and Hidden Layer
w_{jk}	Weight between Hidden and Output Layer
Y_K	Output Layer Neuron
Z_j	Hidden Layer Neuron

Pengenalpastian Imej Virus Biologi Menggunakan Pengelas Kepintaran Buatan

ABSTRAK

Tesis ini menggunakan kepintaran buatan untuk menghasilkan satu sistem bagi mengklasifikasikan jenis-jenis virus biologi. Sebagai permulaan untuk membangunkan sistem ini, sejumlah imej virus biologi yang berlainan jenisnya telah digunakan untuk kerja penyelidikan ini. Dalam tahap pertama, sebelum virus boleh diklasifikasikan, pertama sekali imej virus tersebut perlu di praproses dengan mengenakan pencemaran bunyi yang berlainan tahap terhadap imej tersebut. Ini adalah kerana di dalam bidang pengklasifikasian virus biologi, imej terperinci virus adalah sangat penting memandangkan betapa dekat sesetengah virus menyerupai virus yang lain walaupun berlainan jenis. Jadi, praproses imej adalah penting bagi memastikan pengekstrak ciri dapat mengekstrak ciri dengan tepat dan membolehkan algoritma klasifikasi menghasilkan ketepatan yang tinggi. Proses ini dapat dicapai dengan mengenakan bunyi terhadap imej tersebut. Terdapat berbagai jenis bunyi yang boleh digunakan untuk mencapai objektif ini. Bagi tujuan penyelidikan ini, jenis bunyi *salt&pepper* telah digunakan dalam pembangunan sistem. Untuk tahap kedua, setelah bunyi telah digunakan untuk mencemari imej virus, beberapa jenis pengekstrak ciri telah dibangunkan untuk mengekstrak keluar ciri imej. Pengekstrak ciri memainkan peranan yang penting bagi mendapatkan ketepatan klasifikasi yang tinggi. Tahap ketiga eksperimen ini memerlukan pembangunan beberapa jenis klasifikasi algoritma kecerdasan buatan. Ciri yang telah diekstrak akan dihantar kepada klasifikasi algoritma kecerdasan buatan ini di mana algoritma ini akan menggunakan ciri tersebut untuk melihat betapa tepat imej virus biologi dapat diklasifikasikan. Hasil ketepatan adalah berdasarkan kepada gabungan jenis pengekstrak ciri dan juga algoritma klasifikasi yang digunakan. Hasil eksperimen pada masa sebenar yang dijalankan telah menunjukkan sistem ini dapat menunjukkan kekuatan, ketelitian, dan kecekapan yang tinggi dalam mengkalasifikasikan imej virus biologi di mana ketepatan sehingga 99.93% telah dicapai oleh gabungan pengekstrak ciri dan algoritma klasifikasi yang diusulkan. Kombinasi pengekstrak ciri dan algoritma klasifikasi yang diusulkan telah menghasilkan keputusan yang lebih tinggi dibandingkan dengan kebanyakan aplikasi masa sebenar yang lain.

Biological Viruses Images Recognition Using Artificial Intelligence Classifier

ABSTRACT

This thesis uses Artificial Intelligence for biological viruses images classification. The start of the development of this system requires a number of biological viruses images which belongs to different classes for this research work. The first stage of this system before the virus can be classified are as such, the images first needs to be preprocessed using noise density of different level. It is crucial because in biological viruses images classification, the detail of the viruses images are important due to how closely some of the viruses resembles each other even though they are not in the same classes type of virus. Therefore, image preprocessing is required for ensuring the feature extraction would be able to correctly extract the viruses details allowing the classification algorithm to produce a high accuracy results. This process are done by smearing the images with noise. For this research work, salt&pepper noise are used in the system development. The second stages are after the noise have been applied to the images, several feature extraction methods have been developed to extract the images feature out. Feature extraction method plays an important part in this research for determining and acquiring the best possible results accuracy. The third stage of this experiments required several artificial intelligence classification algorithms. The algorithm will classify and produce the results accuracy based on the extracted feature from feature extraction method. This will show the accuracy of the feature extraction method is based on how accurate the biological viruses are classified. The accuracy is based on the combination of feature extraction method and artificial intelligence classification algorithm. The real time experiment conducted proved that the proposed feature and classifier combination are robust, excellent, and efficient of which it has produced a results accuracy of up to 99.93% for biological viruses images classification. The proposed combination produced a much better result as compared with most of the real time applications of this system.

CHAPTER 1

INTRODUCTION

1.1 Project Background

Virus is commonly known as a type of infective agent which typically consists of protein coated, nucleic acid molecule. To be more precise, virus is actually a microscopic organism which consists of genetic materials such as RNA or DNA, surrounded by either protein, glycoprotein, or lipid (fat) coating. Viruses are often described as microorganisms, however some microbiologists believe viruses are a type of microscopic infective agent, due to it being a non-living organism. Viruses are able to multiply itself but only within the living cells of a host as it is too small to be seen even by the use of light microscopy (Racaniello,2004).

Virus detection is one of the topics that is common in the focus of object detection and image recognition. Virus is one of the most important features of which it can be used to distinguish several types of visual objects. It is well known that learning and detecting viruses is a challenging and complex task even for computer systems (Garces et al., 2016). This is partly due to the lighting and changes in viewpoint on the surface region. Since viruses are known to be too small to be identified even by light microscopy a technique that is needed for detecting these viruses.

In order to address this limitation, negative stain Transmission Electron Microscopy (TEM) is needed as an effective microscopy technique which can produce distinctive surface textures of the viruses. It has been proven to be a very valuable asset especially in virus detection, its discovery, and the taxonomy of the virus (Goldsmith and Miller, 2009). Recently, classification of TEM images has always been exclusively performed with the microscope by expert personnel which is known to be expensive and are not entirely reliable since the results depend heavily on the skills and experiences of each expert inspectors (Rajan, 2013). This has made the automation of negative stain Transmission Electron Microscopy highly desirable.

Virus shapes usually range from icosahedral patterns to highly pleomorphic particles. It's shape and size alone are however insufficient to be used to confirm specific types of viruses. This is where TEM virus images play an important role since it is very well suited for analysis using machine patterns due to their properties of shape, size, and texture. Texture provides indispensable information needed to confirm specific virus types since many viruses show recurring and distinct texture patterns.

In the last couple of decades, much work has been done on computer vision focusing on image textures (Nanni et al.,2013). Early texture classification method is used for exploring the statistical analysis of images. It is to propose the most efficient and effective operator for the purpose of describing the local image pattern and simultaneously produce an impressively accurate image classification result (Xu et al., 2011). This research investigates the performance of pre-processing stages with different combinations of feature extractions and classifiers. This will allow the most effective and efficient combinations of feature extractions and classifiers to be identified for this project.

The virus images need to be pre-processed. Different types of feature extractions will be employed which include the Principal Component Analysis (PCA), Local Binary Pattern (LBP), and Gray Level Co-Occurrence Matrix (GLCM). The results gathered from extracted features are then employed to artificial intelligence classifiers such as Neural Network (NN), *k* -Nearest Neighbor (*k*-NN), Fuzzy *k*NN (*F*-*k*NN), and Linear Discriminant Analysis (LDA). *k*-Nearest Neighbor algorithm (*k*-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression. Fuzzy *k*-Nearest Neighbor (*F*-*k*NN) algorithm is utilized to develop a fuzzy wave height prediction model for large lakes, where the fetch length depends on the wind direction.

Artificial Intelligence is universal and also a highly flexible function used in the cognitive and engineering sciences field. In this research, to classify the viruses, it is best proposed to employ the artificial intelligence method. A MATLAB program will be developed for the purpose of extracting the features and developing the artificial intelligence algorithm classifiers. From then, by using suitable feature extraction techniques, the feature vectors are derived. The extracted feature details will be used for the artificial intelligence method. The developed system will be tested for results validation.

1.2 Problem Statement

As mentioned in the previous section, learning and detecting virus is a challenging and complex task for computer systems. This is due to the lighting and viewpoint changes on the surface region. This hinders the research efforts of

recognizing and classifying virus types. Without negative stain Transmission Electron Microscopy (TEM), which allows the shape, size and texture of the virus to be analysed, it will be insufficient to confirm specific virus types by depending only on the virus shape and size alone. Despite the recent advancement in the image texture analysis field and the crucial role of texture in TEM virus classification, there are few papers and researches done that examines machine analysis of virus textures in transmission electron microscopy images (Nanni et al.,2013).

Even though as of late, texture and image recognition have been booming, there is a lack in commercial technologies under the generic topic of virus image recognition and classification because of how challenging and complex it is for computer systems to recognize and classify virus types and how it is insufficient to classify a virus depending solely on the virus shape and size alone to correctly confirm it, making it harder for classification (Nanni et al.,2013).

There are a number of sources of variations to classify a virus as mentioned in the above, such as a virus's shape, size, and texture which in turn will be helpful in extracting the virus's image features and will then be used by artificial intelligence classifiers to classify the virus according to its type. This variation is called intrinsic source of variation. The extrinsic source of variation is an imaging process which includes, focus, resolution and noise (César & Sang,2011).

The extrinsic source of variation may or may not prove to hinder the recognition process. This depends on the algorithm that the feature extraction method uses. It is however possible that the variation influenced by factors, such as noise, makes a difference in classifying the virus. However, as said before, a lot of work has been poured into computer vision, focusing on image textures, within the last one to two

decades. Thus, this research seeks to find the most efficient feature extractions under different variants of salt and pepper noise, for biological virus image recognition.

1.3 Objectives

The objectives of this research are as stated:

1) To observe effect of salt and pepper noise density levels for biological virus's image recognition

2) To find important features from the virus's database

3) To determine the performance of virus recognition systems with various classifiers.

1.4 Scope

This work is only limited to five difficult viruses that cannot be distinguished such as Adenovirus, Astrovirus, Cowpox, Dengue and Ebola from the TEM virus database. The database of whole virus images in this work only consists of upfront images and does not deal with different poses. Within this work, only common features such as PCA, LBP and GLCM are adopted. However, different parameters and coefficient of features under different noises are acquired. Salt and pepper noise is used instead of other noises because this type of noise always appear in digital images and often times appropriated as a benchmark for filter performance evaluations

1.5 Research framework

This research employs supervised learning for classification where the inputs are categorized into five classes, and the learner assigns these inputs according to each member of the classes.

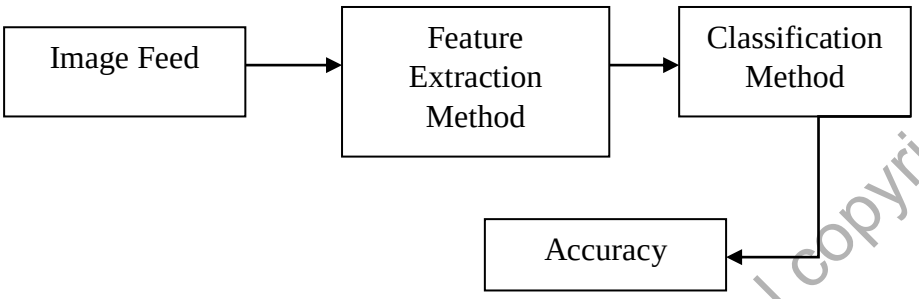

Figure 1.1: Supervised Machine Learning

1.6 Thesis Organization

Chapter 1 is the introduction of the thesis where the project background is discussed. The objectives of the research and scope will also be discussed in this chapter. The focus of this project is mainly about learning and detecting virus by using the computer system

In Chapter 2, there will be a brief explanation about the image pre-processing method and the feature extractions method. The artificial intelligence classifiers that will be discussed are Neural Network (NN), *k*-Nearest Neighbor (*k*NN), Fuzzy *k*-Nearest Neighbor (*F*-*k*NN) and Linear Discriminant Analysis (LDA).

Chapter 3 is about the design and development of image pre-processing and feature extraction methods of biological virus image recognition system. There are three different feature extraction methods used throughout this entire research which are the Principal Component Analysis, Local Binary Pattern and Gray Level Co-Occurrence

Matrix. Local Binary Pattern is used to describe the local image pattern. Basically, it is a type of visual descriptor commonly used in the classification of computer vision.

Chapter 4 will then provide an in-depth explanation about the classifiers for biological viruses' image. Every classifier will be discussed in detail and the algorithms involved will also be presented. The performance analysis and the results of the experiments will be introduced in Chapter 5. Lastly, Chapter 6 will contain a discussion about future works and conclusions of this project.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

As mentioned in the introduction, the classification process of TEM images is exclusively performed under the microscope by expert personals. This proved to be a problem since the reliability of the results depend highly on the experiences and the skills of each expert observer, and the observation itself requires a hefty amount of expenses to be done. Therefore, other methods for the development of TEM image classifications of biological viruses should be taken into consideration

In general, the decision-making process of a human mind might be based off an unconscious intuition or experience which even if disputed and refuted, the decision that has been made will not change. This process usually takes an entire situation into account thus making a specification into isolated observations and arguments become much more difficult (César & Sang, 2011). This reasoning raises the exact question needed to be asked which is whether it is possible to constitute the method and develop a fully automatic pattern recognition using machine analysis method instead.

Supposedly it is possible for a virus to be recognized. Can the virus be named and does determining the virus type or class of which it belongs to possible? It is undeniable that some representation is needed if the virus images are to be fed to a computer for an automatic recognition. Viruses differ from each other. It is impossible to have an exact match unless it is the same exact virus. Often times, even though the

viruses are of the same type, there is a slight difference in their features making it challenging for image or texture recognition. Can the viruses be classified and identified if it is of the same type since the difference between every similar type virus is almost non-existent? This is therefore the purpose of conducting this research which is to find the most efficient feature extractions and artificial intelligence classifiers for biological virus image recognitions in order to get the most accurate results.

However, image classification has always proven to be a complex process since results may vary based on multiple different factors. This literature review suggests that by using multiple features and a number of selected suitable classification methods are highly significant for producing and improving the accuracy of biological virus image recognition. Non-parametric classifier such as neural network, fuzzy, and knowledge-based classification nowadays have steadily and increasingly become a rather important approach for the purpose of classification. (Lu & Weng, 2007).

This chapter will focus on major methods involved in biological virus image recognition and classification.

a) Template-based approach is most effective when used on a bulk of template image or for a template with strong features.

b) Feature-based approach is perfect for an image that has a strong feature, which may prove to be further useful if the searched image transforms in any way.

c) Knowledge-based method is an approach which reasons and uses knowledge as a base to classify the extracted feature of viruses.

Some of the major steps in classification includes suitable classification systems, the selection of training samples and image pre-processing, extracting features from an image, a selection of suitable classification algorithms, and also accuracy assessment.

(Lu & Weng, 2007). A brief survey of how these steps are implemented in this research is presented in this chapter.

As mentioned above, classifying and recognizing virus based on images have always been proven to be a challenge because of how identical it is if it belongs to the same class. Therefore, the first and most important step of this research is to find a selection of training samples. The samples are then pre-processed to determine which condition allows the program to extract the viruses' features with the highest accuracy. The accuracy assessment is gathered using suitable classification approaches. In this case, all the images are pre-processed before the features are extracted. This research uses Principal Component Analysis (PCA), Gray-Level Co-occurrence Matrix (GLCM), and Local Binary Pattern (LBP) for feature extractions. The results produced are then written down and recorded. The extracted feature will then be used by the artificial intelligence algorithm classifiers which is Neural Network (NN), *k* Nearest-Neighbor (*k*NN), Fuzzy *k*-Nearest Neighbor (*F*-*k*NN) and Local Discriminant Analysis (LDA) classifiers. The accuracy results gathered by using the combination of different feature extraction methods and artificial intelligence classifiers are later discussed in chapter 5.

2.2 Image Pre-processing Method

Pre-processing operations which are also referred to as image restoration and rectification are intended for sensor correction or geometric distortion of data. It is valued for its use in data acquisition and its conditions during the data acquisition process. In this research, it is desirable to convert and calibrate the data which is the feature extracted from the image of the viruses to facilitate comparison between the data (Barni, 1998).