



UNIVERSITI PUTRA MALAYSIA

**PERSONAL IDENTIFICATION BY KEYSTROKE PATTERN FOR
LOGIN SECURITY**

NORHAYATI BT ABDULLAH

FSKTM 2001 1

**PERSONAL IDENTIFICATION BY KEYSTROKE PATTERN FOR
LOGIN SECURITY**

By

NORHAYATI BT ABDULLAH

**Thesis Submitted in Fulfilment of the Requirement for the Degree of Master of
Science in Faculty of Computer Science and Information Technology
Universiti Putra Malaysia**

August 2001



This book is dedicated to my children Aizat and Aqil in the hope that it will give them inspiration and courage to achieve as high as they can in their education.

Remember,

*Education is difficult and expensive. But whatever it costs,
it's cheaper than ignorance.*

*May the Blessings of Allah
be upon them.*

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master Science

**PERSONAL IDENTIFICATION BY KEYSTROKE PATTERN FOR
LOGIN SECURITY**

By

NORHAYATI BT ABDULLAH

August 2001

Chairman: Ramlan Mahmud, Ph.D.

Faculty : Computer Science and Information Technology

This thesis discusses the Neural Network (NN) approach in identifying personnel through keystroke behavior in the login session. The keystroke rhythm that falls in the behavioral biometric has a unique pattern for each individual. Therefore, these heterogeneous data obtained from normal behavior users can be used to detect intruders in a computer system.

The keystroke behavior was captured in the form of time within the duration between the pressing and releasing of key was recorded during the login session. Ten frequent loggers were chosen for the experiments. The data obtained were presented to NN for pattern learning and classifying the strings of characters. The backpropagation (BP) model was implemented to identify the keystroke patterns for each class.



Various architectures were employed in the BP training to achieve the best recognition rate. Several features that influence the network were considered. The experiment involved the slicing of input data and the determination of the number of hidden units. Several other factors such as momentum, learning rate and various weight initialization were used for comparison. Three types of weight initialization were used, including Nguyen-Widrow (NW), Random and Genetic Algorithm (GA). The experiment showed that the recognition of 97% was achieved using NW weight initialization with 10 hidden units. Further experiments with Improved Error Function (IEF) in standard BP has showed better results with 100% recognition on both train and test data set compared to previous experiment.

The results of this study were compared with Chambers's (1990) and Obaidat's (1994) work. Chambers used the data set similar to the data used in this experiment and obtained 90.5% recognition through Inductive Learning Classifier method, while Obaidat used standard BP with 6 classes and obtained 97.5% recognition.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk Ijazah Master Sains

**MENGENAL PERSONAL MELALUI CORAK TEKANAN PAPAN KEKUNCI
BAGI KESELAMATAN LOG MASUK**

Oleh

NORHAYATI BT ABDULLAH

Ogos 2001

Pengerusi: Ramlan Mahmud, Ph.D.

Fakulti : Sains Komputer dan Teknologi Maklumat

Tesis ini membincangkan pendekatan Rangkaian Neural (RN) dalam mengenalpasti personal melalui perlakuan tekanan kekunci semasa sesi log masuk. Rentak tekanan kekunci yang dikategorikan di dalam biometrik perlakuan mempunyai corak unik untuk setiap individu. Oleh itu, data heterogenus ini yang diperolehi daripada pengguna berpelakuan normal boleh digunakan untuk mengesan pencerobohan di dalam sistem komputer.

Perlakuan tekanan kekunci diperolehi dalam bentuk masa di mana tempoh antara pengguna tekan kekunci dan lepas direkodkan semasa sesi log masuk. Sepuluh pengguna yang bekerapan log masuk dipilih dalam eksperimen. Data yang diperolehi diberi kepada RN untuk pembelajaran corak dan pengelasan rentetan aksara. Model rambatan balik (BP) dilaksanakan untuk mengecam corak tekanan kekunci untuk setiap kelas.

Pelbagai rekabentuk digunakan dalam latihan BP untuk mencapai kadar pencaman terbaik. Beberapa fitur yang mempengaruhi rangkaian telah dipertimbangkan. Eksperimen ini melibatkan cincangan data input dan penentuan bilangan unit tersembunyi. Beberapa faktor lain seperti momentum, kadar pembelajaran dan pengistiharan awal pemberat telah digunakan iaitu Nguyen-Widrow (NW), Rawak dan Algoritma Genetik (AG). Eksperimen ini menunjukkan pencaman sebanyak 97% telah dicapai menggunakan NW dengan 10 unit tersembunyi. Eksperimen selanjutnya yang menggunakan kaedah Pembaikan Semula Fungsi Ralat (PSFR) di dalam BP piawai telah menunjukkan keputusan yang lebih baik dengan kadar pencaman 100% ke atas kedua-dua set data latihan dan data ujian berbanding dengan eksperimen sebelumnya.

Keputusan daripada kerja ini telah dibuat perbandingan dengan kerja Chambers (1990) dan Obaidat (1994). Chamber menggunakan set data yang sama dalam eksperimen ini dan memperolehi 90.5% pencaman melalui kaedah Pengelas Pembelajaran Induktif, manakala Obaidat menggunakan BP piawai dengan 6 kelas memperolehi 97.5% pencaman.

ACKNOWLEDGEMENTS

In the name of Allah – Most Merciful, Most Compassionate

First of all, I would like to express my gratitude to my supervisory committee chaired by Dr. Ramlan Mahmud, the committee members, Dr. Md. Nasir Sulaiman and En. Razali Yaakob for their helpful guides, comments and suggestions during my study here. They have given me fruitful knowledge and experience in my research work. I would also like to thank Dr. Hamidah Ibrahim, Faculty's Co-ordinator of Graduate Studies for providing the hardware and comfortable lab to work in.

My appreciation also goes to Mr. J.A. Michael Chambers, Chartered Information Systems Practitioner of AmpsToll Incorporated, New York for his effort in delivering the keystroke data and kind guidance in my early work.

To my dear colleagues, Siti, Ummu, Ija, Iza, Anom, Umi, Lay Ki, and the rest, thank you for being supportive. I also would like to thank personally to En. Saliman Manaf of Mimos Bhd. for helping me to configured the Linux PC. Not forgetting the faculty technical support team, thank you for your technical support.

To my mother, Nik Selamah bt Wan Mohd.Noor, dear husband and children, and family members, thank you for your firm support.



Last but not least, I would like to thank the Public Service Department for sponsoring my study in Universiti Putra Malaysia.

Wassalam.

I certify that an Examination Committee met on 7th August 2001 to conduct the final examination of Norhayati Abdullah on her Master of Science thesis entitled “Personal Identification by Keystroke Pattern for Login Security” in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

ALI MAMAT, Ph.D.

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

RAMLAN MAHMUD, Ph.D.

Deputy Dean
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

MD. NASIR SULAIMAN, Ph.D.

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

RAZALI YAAKOB

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)



MOHD GHAZALI MOHAYIDIN, Ph.D.
Professor/Deputy Dean of Graduate School,
Universiti Putra Malaysia

Date 27 OCT 2001



This thesis submitted to the Senate of Universiti Putra Malaysia has been accepted as fulfilment of the requirement for the degree of Master of Science.



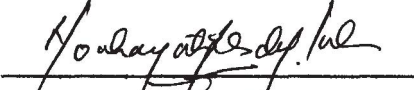
AINI IDERIS, Ph.D.
Professor,
Dean of Graduate School
Universiti Putra Malaysia

Date:

13 SEP 2001

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.



NORHAYATI ABDULLAH

Date: 27/05/2001

TABLE OF CONTENTS

	Page
DEDICATION	ii
ABSTRACT	iii
ABSTRAK	v
ACKNOWLEDGEMENTS	vii
APPROVAL SHEETS	ix
DECLARATION FORM	xi
LIST OF TABLES	xiv
LIST OF FIGURES	xv
LIST OF ABBREVIATIONS	xvii
CHAPTER	
I INTRODUCTION	1
Introduction	1
Problem Statement	2
Objective	3
Scope of Work	3
Organization of the Thesis	4
II LITERATURE REVIEW	5
Introduction	5
Computer Security	5
Intrusion Detection	6
Biometric	11
Keystroke	15
Scan Codes	16
Keystroke for Identification	17
Neural Network	20
BP in Biometric Applications	23
NN for Keystroke Identification	23
III SYSTEM ARCHITECTURE	25
Introduction	25
System Architecture	25
Pre-Processing Module	26
Data Preparation	26
Neural Network Module	33
BP Model	34



	Learning Rules	35
	Training Phase	36
	Recognition Phase	41
	IEF of Standard BP Model	42
	Weight Initialization	42
	Random Weight Initialization	43
	NW Weight Initialization	44
	Genetic Algorithm	45
	Output Module	47
	Hardware and Software	47
IV	RESULT AND DISCUSSION	48
	Introduction	48
	Early Experiments with Binary Inputs	49
	Experiments with Number of Hidden Units	49
	Experiments with Momentum and Learning-Rate	50
	Various Weight Initializations	52
	Standard BP with IEF	54
	Comparison Results of Previous Work	57
	Discussion	57
V	CONCLUSION	60
	Conclusion	60
	Future Work	61
	BIBLIOGRAPHY	62
	APPENDICES	
A	Keystroke Data	66
B	Input File Structure	76
	BIODATA	79



LIST OF TABLES

Tables	Page
2.1 Keyboard scan codes	16
3.1 Raw data of single login session.....	27
3.2 Scan code representation per user	28
3.3 Sequence of keystroke per user	28
3.4 Refined data of single user.....	29
3.5 Example of a set of users keystroke	36
3.6 Target file.....	37
4.1 Input data with 3 decimal points	51
4.2 Input data with 4 decimal points	51
4.3 Experiment results with NW weight initialization	54
4.4 Experiment results with random weight initialization	54
4.5 Experiment results with GA weight initialization	54
4.6 Experiment results of various Algorithm.....	55
4.7 Result of Standard BP with IEF	56
4.8 Experiment results with various data set using standard BP...	57
4.9 Experiment results with 10 classes using multiple classifiers	57



LIST OF FIGURES

Figures		Page
1	Intrusion Statistics in Malaysia	9
2	A Block Diagram of Typical Anomaly Detection System	10
3	A Block Diagram of Typical Misuse Detection System	11
4	The Bertillon System	13
5	AT Keyboard Layout with Scan Code	15
6	Biological Neuron	21
7	Neural Network Model Proposed by MC Culloch and Pitts	21
8	General NN System Architecture.....	25
9	Pre-Processing Module.....	26
10	Keystroke Pattern of User-Id “Abevan”.....	29
11	Keystroke Pattern of User-Id “Amcarrin”.....	29
12	Keystroke Pattern of User-Id “Aarmstro”.....	30
13	Keystroke Pattern of User-Id “Gyen”.....	30
14	Keystroke Pattern of User-Id “Jalesper”.....	30
15	Keystroke Pattern of User-Id “Jew”.....	31
16	Keystroke Pattern of User-Id “Mbuehner”.....	31
17	Keystroke Pattern of User-Id “Schao”.....	31

18	Keystroke Pattern of User-Id “Sjette”.....	32
19	Keystroke Pattern of User-Id “Wtpchui”.....	32
20	Multilayer NN Model.....	33
21	BP Network Structure	35
22	Input File	37
23	Example of Output from the Recognition Phase	41
24	Various Hidden Units with $\alpha = 0.9$ and $\beta = 0.02$...	49
25	BP with Various Inputs	51
26	Generalization Effect in BP Learning	52
27	Recognition Rate Vs Initial Weights	54



LIST OF ABBREVIATIONS

ASCII	-	American Standard Code for Information Interchange
BIOS	-	Basic Input Output System
BP	-	Backpropagation
CERT/CC	-	Computer Security Incident Response Teams / Coordination Center
GA	-	Genetic Algorithm
ID	-	Intrusion Detection
IDES	-	Intrusion Detection Expert System
IEF	-	Improve Error Function
MLP	-	Multi-Layer Perceptron
MYCERT	-	Malaysia Computer Security Incident Response Teams
NIPS	-	Northern Island Prison Service
NN	-	Neural Network
NW	-	Nguyen-Widrow
OS	-	Operating System
PC	-	Personal Computer
SPAWAR	-	Navy's Space and Naval Warfare System Command
WATSAR	-	Waterloo Student Workstation



CHAPTER 1

INTRODUCTION

Introduction

Prevention against unauthorized users from accessing information in any system is the first element of defense against intruders. A system must first identify a user to determine access privileges and track what the user does. This implies that there must be unique identifiers for all users. A system must also authenticate users, that is, to verify that they are who they say they are. These two tasks are combined into one mechanism, which is called the login process.

There are several ways of detecting unauthorized users attempting to invade the system. One way is through intrusion detection systems. Intrusion detection (ID) is the process of monitoring activity on a system in real time for the purpose of identifying attempts or successful intrusion of the system. Artificial intelligence (AI) techniques such as data reduction and classification have been used in many ID systems (Frank, 1994). The statistical approach has also resulted in systems being used and tested extensively in ID system (Kumar and Spafford, 1994).

Neural Network (NN) is a classifier system that uses a model biological system to perform classification. Because of its ability to learn and generalize, NN has been widely used in many applications such as pattern mapping and classification. Human behavior based on keystroke characteristics are vary from person to person. The way a person depresses a key produces a different timing from other people. NN is

trained with the timing vectors of the owner's keystroke rhythm to discriminate between the owner and an imposter. Implementing of NN in keystroke application has shown remarkable results of recognition.

Problem Statement

Often accessing a system requires some unique identification. Everyone has characteristics that make him or her unique. Keystrokes for example, are individual patterns and rhythms of typing repetitive character groups. A true user keys his or her login name more consistently than does a forger. A forgery might be good in impersonating but as mentioned before, the way he or she keys in can be detected through rhythms of typing.

Recently, Obaidat (1994), applied classical pattern recognition techniques to the individual's typing technique to achieve user identification. Joyce and Gupta (1990) have described their method of using keyboard latency information captured during a user's login process. John A. Robinson et al. (1998) reported that an application of typing style analyses of very short strings (login names) has given insights into typing style identification through keystroke dynamics.

Experiments have been made using multi-layer NN to identify users. In a research by Obaidat, three types of networks were used; BP, Sum-of-Product (SOP) and a new hybrid architecture that combines both (Obaidat,1994). The experiment with SOP network did not seem practical for this problem. SOP network has taken up the real training time due to a large number of hidden units a for small input units. On the other hand, the Hybrid SOP gained a better recognition with only 5 hidden units.

Obviously, the number of hidden units being applied within 4 to 5 units to BP was not sufficient for the internal processing during training.

Bleha (1993) has used the perceptron algorithm on some simple applications to verify the identity of computer users with fairly good results. The perceptron is a learning device. In its initial configuration, the perceptron was incapable of distinguishing the patterns of interest through a training process but it could achieve the capability under certain conditions (Freeman et al., 1991).

To reduce some of the problems faced by the previous work, this study works on multi-layer neural network with various parameters like the initial weight, hidden units, momentums and learning rate. It is hoped that by experimenting with various architectures, the findings will contribute to better work in this area.

Objective

The objective of this work is to identify computer users by their keystroke pattern.

Scope of Work

NN model is being implemented in keystroke pattern recognition. This work will include 10 users' login identification (users-id) that the network can recognize, with each user keying in 20 times.

For comparison, three types of weight initialization: Nguyen Widrow, random, and genetic algorithm were used. Beside the various weight initializations, other factors

such as various hidden units, momentum, learning rate and Improved Error Function (Shamsuddin et al., 2001) were used for comparison. The input presented in bipolar may improve the network learning (Fausett, 1994). Therefore, bipolar sigmoid activation functions were used in the experiment.

Organization Of The Thesis

The thesis is organized as follows. Chapter II addresses the historical perspective of identity verification followed by literature review on the ID, biometric, keystroke operation, and available tools to monitor keystroke rhythms for ID, a brief review on NN and implementation of NN in keystroke identification. Some previous works on keystroke pattern recognition were also mentioned in this chapter.

Chapter III focuses on the system architecture. This chapter outlines the processing stages of data before data can be used later for experiments. Further discussions on BP and various type of architectures are also included here.

Chapter IV describes the experimental work on keystroke recognition in NN, the result and discussion. Various architectures have been applied in the experiment. This chapter also describes experiments with IEF in standard BP and the results were compared with previous works on keystroke.

A summary and conclusion is contained in Chapter V. There are also suggestions for future work that may extend the use of NN model in keystroke recognition.

CHAPTER II

LITERATURE REVIEW

Introduction

In this chapter we shall illustrate the need for securing computer systems, provide a history of intrusion detection plus an analyses of incidents that occurred, as well as techniques and definitions of each subject. An overview in the field of biometric use, keystroke and a brief discussion on the emergence of neural network to the latest trend of neural net in network security are also presented in this chapter.

Computer Security

Computer security was of little concern in the early days of computing. The number of computers and the number of people with access to those computers was limited. The first computer security problems, however, emerged as early as the 1950's, when computers began to be used for classified information (Howard, 1997). Confidentiality (also termed secrecy) was the primary security concern, and the primary threats were espionage and the invasion of privacy. At that time, and up until recently, computer security was primarily a military problem, which was viewed as essentially being synonymous with information security. From this perspective, security is obtained by protecting the information itself. By the late 1960's, the sharing of computer resources and information, both within a computer and across networks, presented additional security problems. Computer systems with multiple users required operating systems that could keep users from intentionally or inadvertently interfering with each other. Network connections also provided



additional potential avenues of attack that could not generally be secured physically. Towards the millennium, computer security has become the first issue to the connected world.

A narrower definition of computer security is based on realization of confidentiality, integrity, and availability in computer systems (Russel & Gangemi, 1991). Confidentiality requires that information be accessible only to those authorized for it; integrity requires that information remain unaltered by accidents or malicious attempts, and availability means that the computer system remains working without degradation of access and provides resources to authorized users when they need it. By this definition, unreliable computer systems are unsecured if availability is part of its security requirements.

Identification shall be defined as consisting of those procedures and mechanisms that allow agents external to some computer system to notify the system of their identity (Amoroso, 1994). The need to perform identification techniques arises when one wishes to associate each action with some agent that causes each action to occur. Practical computer systems can determine who invoked an operation by examining the reported identity of the agent who initiated the session in which that operation is invoked. This identity is most typically established via a login sequence.

Intrusion Detection

Intrusion detection is the process of monitoring the events occurring in computer systems or networks, analyzing them for signs or security problems (Bace, 2000). Its



research and development only emerged progressively around the 1980's. Funded by the U.S Navy's Space and Naval Warfare System Command (SPAWARS), Dorothy Denning and Peter Nuemann (from 1984 to 1986) researched and developed a model for real-time intrusion detection system, named the Intrusion Detection Expert System (IDES). This research proposed a correlation between anomalous activity and misuse. Within the same period, SPAWARS also funded another project called Automated Audit Analysis. This research demonstrated the capability to distinguish normal from abnormal usage. Another expert system called Discovery used statistical inference to locate patterns in the data input. The system was designed to detect three types of abuse scenarios such as unauthorized access, insider misuse, and invalid transactions. Until the 1990's, intrusion detection systems were largely host-based, confining their examination of activity to operating system audit trails or host-centric information sources.

As a society we are becoming increasingly dependent on rapid access and processing of information. Increased connectivity not only provides access to larger and varied resources of data more quickly than ever before, it also provides an access path to the data from virtually anywhere on the network. Consequently this may lead to many computer leakages, intrusions, attacks and many more terms that are referred to as computer crimes. Computer viruses are the most common and well-known attack against computers. An attack is a single unauthorized access attempt, or unauthorized use attempt, regardless of success. On the other hand, an incident involves a group of attacks that can be distinguished from other incidents because of the distinctiveness of the attackers, and the degree of similarity of sites, techniques, and timing. An attack can be categorized into seven types as follows:

