OFFLINE HANDWRITTEN ARABIC WRITER IDENTIFICATION USING NEGATIVE SELECTION ALGORITHM

HAGER A. SULIMAN

UNIVERSITI TEKNOLOGI MALAYSIA

OFFLINE HANDWRITTEN ARABIC WRITER IDENTIFICATION USING NEGATIVE SELECTION ALGORITHM

HAGER A. SULIMAN

A dissertation report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

JULY 2014

To my husband, Ahmed, who has been a constant source of support and encouragement during the challenges of study and life

To my Mother, Fatema, who has been a source of motivation and strength during moments of despair and discouragement

My children, Mohamed, Hamzah and Ibrahim

I give my deepest expression of love and appreciation for the encouragement that they gave and the sacrifices they made during this graduate program

.....

.....

.....

I dedicate this Second fruit ... From Malaysia

ABSTRACT

In pattern recognition; writer identification is one of the research are as that attract the researchers' interest in the conduct of their studies. Writer's identification of identity and its determination ability is not the only important thing to the writer, but the accuracy of this determination is considered as a big challenge. This study evaluates the accuracy of Arabic Handwriting Identification performance using the Bio-Inspired classifier. The study shows that the accuracy of the identification performance could be greatly improved with the Bio-Inspired classifier. The framework of the writer identification consists of three main phases: pre-processing phase, feature extraction phase, and classification phase. This research adopts IFN/ENIT Arabic Database which is constructed by Ecole National ed'Ingénieur de Tunis (ENIT) in Tunisia and Institute of Communications Technology in Germany (IFN). The images are enhanced by applying the threshold and conversion of the grayscale level images into black and white. Geometric Moment Function is used to extract the features from the images. Finally, the Bio-Inspired classifier is applied in this research with the use of Negative Selection Algorithm to classify and identify the writer. The obtained results show a promising ability of NSA in Writer identification. Other researchers could apply the NSA on handwriting languages that uses the same Arabic letters with different semantic such as Urdu as well as Farsi.

ABSTRAK

Pengecaman Paten merupakan sebuah bidang yang luas yang mana di antaranya adalah merupakan bidang Pengenalpastian Penulis yang kini semakin popular serta berjaya menarik minat penyelidik dalam menjalankan kajian penyelidikan mereka. Menerusi bidang Pengenalpastian Penulis, ketapatan penentuan merupakan cabaran yang besar selain daripada cabaran lain seperti cabaran pengenalpastian identiti penulis dan cabaran keupayaan penentuan. Kajian penyelidikan ini menilai ketepatan prestasi terhadap Pengenalpastian Tulisan Tangan Bahasa Arab melaui cara pengkelasan bioinspirasi. Melalui kajian pengkelasan bioinspirasi yang dijalankan ini, prestasi pengenalpastian menunjukan penambahbaikan yang boleh diperolehi. Rangka kerja pengenalpastian penulis itu terdiri daripada tiga fasa utama iaitu fasa pra-pemprosesan, fasa pengekstrakan ciri, dan fasa pengkelasan. Penyelidikan ini menggunakan IFN/ENIT Arabic Pangkalan Data yang dibina oleh Ecole Nationale d'Monthly de Tunis (ENIT) di Tunisia dan Institut Teknologi Komunikasi dalam Jerman (IFN). Imej-imej dipertingkatkan dengan menggunakan nilai ambang dan penukaran imej tahap skala kelabu ke hitam dan putih. Geometri Moment Fungsi digunakan untuk mengekstrak ciri daripada imej-imej. Akhir sekali, pengkelasan bioinspirasi digunakan dalam kajian ini bersama dengan penggunaan Algoritma Pemilihan Negatif bagi tujuan pengekelasan dan pengenalpastian penulis. Keputusan yang diperolehi menunjukkan keupayaan NSA yang berpotensi dalam pengenalpastian Penulis. Penyelidik lain boleh menggunakan NSA pada bahasa tulisan tangan yang menggunakan huruf Arab yang sama dengan semantik yang berbeza seperti Bahasa Urdu dan juga Farsi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Pattern Recognition is a branch of Artificial intelligence that deals with the operation and design of systems that recognize patterns of data, these patterns might be a human face; handwritten cursive word, speech signal, finger print image, or a bar code (Parzes and Mahmoud, 2012). As a result of the importance of pattern recognition in various emerging applications such as retrieval and organization of multimedia data bases, data mining, document classification, and biometric authentication (Jain et al.,2004); the interest in this field has been on the increase. There are two kinds of biometric features; these include behavioural (signature, handwriting and voice) and physiological (Iris, face and finger print).

Biometrics can be an effective component for person identification solutions, instead of passwords (which are used in electronic access control) or cards (which are used in banking applications) (Jain *et al.*, 2004). Biometrics has the potential to identify a person uniquely. Thus, it could be employed in detecting crime, identifying criminals, and eliminating fraud. Handwriting identification is considered as a behavioural biometric approach, which is used for identification of people. Virtually, it is how to determine and identify the writer of a handwriting sample between a set of writers. In other words, Writer Identification (WI) is considered as one to-many search in a handwriting database with the return of a list of likely candidates.

Writer identification is still performing manually by forensic handwriting experts. To identify the writer; they observe, compare and evaluate the different features in order to recognize similar unique features from the questioned handwriting in contrast with the original handwriting. The methods of Writer identification are categorized into two kinds; namely: text dependent and text-independent methods. These methods are performed on-line or off-line. Off-line writer identification is considered as the most challenging problem, because of the variability of the writings, and the dynamic information related to the handwriting forms does not exist.

1.2 Problem Background

Hand writing style, signature, fingerprint, iris and face are unique features for humankind, which differentiate one person from the other; this uniqueness makes it possible to establish a handwriting analysis discipline which is essential to criminal justice system and forensic analysis. The challenge in handwriting identification is not only on the ability to determine the writer, but the accuracy of the determination is considered as a big challenge.

Bio-inspired concept uses human immune system concept under soft computing. Human immune system is an adaptive system that has the property of learning. This system can employ so many mechanisms in a parallel way. These mechanisms are implemented for attack of foreign pathogens. The system is able to learn in order to identify and remember what it learnt. Negative Selection Algorithm (NSA) is a Bio- inspired technique which is used in this study to evaluate the accuracy of Arabic handwriting identification performance, and verify whether the accuracy of the identification performance could be greatly improved by the NSA.

The Artificial Immune Systems (AIS) is known has one of the recent biologically inspired approaches that emerge from computer science field (Nemmour

and Chibani, 2013). With the AIS computational technique; many complex problems are solved with the improvement of the computational tools. Pattern recognition, fault detection, classifications, computer security, and optimization are some examples of these problems (Muda *et al.,* 2006; Nemmour and Chibani, 2013).

However, in this research, Negative Selection Algorithm will be used to transform and represent data sets into Bio-Inspired Format. Negative Selection Algorithm (NSA) technique is based on the property of self/non-self to detect foreign antigens; which inspired from the same property in human immune system (Forrest *et al.,* 1994). NSA has three main stages: the first stage is to control the detectors which is generate randomly. The second stage is to monitor the changes by using generated detectors. The last stage is to match the detector set with the new antigens based on matching rule. The Bio-inspired approach in writer identification has three main tasks, these are Granular Data Collector, Bio-inspired Training Environment and Bio-inspired Classifier (Keijzers *et al.,* 2013).

The majority of writer identification researches has been done on English language, where as a few studies has been done on Arabic language although it is spoken by around million people 234 all around the world (Obied, 2010), as well as characters of Arabic language are used to write Arabic, Urdu, Farsi (Persian) languages. Two interrelated problem could be the reason behind this deficiency: the technical challenges and the lack of some required infrastructure to support the writer identification systems in the development.

Arabic language script is different from the English one. First, the writing of the Arabic script is from right to left, also it has 28 characters, each character has four shapes depending on the position of the character within the words. In addition, it contains diacritics which are additional marking controls for the pronunciation of the words. Arabic language has a cursive letters, these letters joined together within a word and that would impact the process of character-level segmentation. All these properties prompt difficulties and challenges in recognizing the Arabic script.

For the identification purpose a standard database should be available. The lack of Arabic databases could be another difficulty in Arabic writer identification researches. Three Arabic databases were constructed for the recognition and identification purposes: Arabic cheque database, Arabic Handwritten Database (AHDB) and IFN/ENIT Arabic database for Tunisian village/town names (Obaid, 2010). However, most of the studies in Arabic handwritten collect and conduct their own dataset that contains number of writers and samples for each writer.

None of the previous work on Arabic writer identification used the NSA as a bio-inspired classifier in the classification phase. Hence, in this study will implements NSA on Arabic hand writing, and evaluate sits effect on the identification performance accuracy using the IFN/ENIT Arabic DB.

1.3 Problem Statement

In handwriting identification all the previous experiments which carried out on AIS classifier on Arabic handwriting have not been done by using the NSA. AIS classifier has the potential of improving the accuracy of writer identification performance. In this study, the AIS performance on the Arabic handwriting will be evaluated and tested using the IFN/ENIT Arabic Database. The difficulties in Handwritten Arabic writer identification related to language barrier (distinct Arabic characters properties) and infrastructure barrier (lack of Arabic Database).

1.4 Project Aim

This project implements AIS computational technique on Arabic handwriting, and evaluates its effect on the identification performance accuracy.

1.4 Objectives

Few objectives have been identified in this study:

i. To apply NSA for off-line Arabic writer identification.
ii. To evaluate the performance of the NSA on handwritten Arabic writer identification.

1.5 Project Scope

i. NSA process is tested on only the Arabic handwriting.
ii. Negative Selection Algorithm is used to transform and represent datasets into Bio- Inspired Format.
iii. The Arabic database which is used is the IFN/ENIT which involves hand written Arabic Tunisian (town/village) names.
iv. Geometric Moment Invariant (GMI) technique is used as afeature extraction.
v. Euclidean distance technique is used for matching between detectors and antigens.
vi. All the algorithms are a performed using Delphi 5 programming language.

1.6 Significance of Project

The lack of research on handwritten Arabic writer identification has motivated the researchers to enrich the area of Arabic handwriting with new untested method; this research proposed implementation of NSA to enhance the accuracy performance of Arabic writer identification.

1.7 Summary

This chapter presents the introduction of the research study, followed by the background and statement of the problem; also it reveals the objectives and scope of the project.

REFERENCES

Abdi, M. N., Khemakhem, M. (2012). Arabic Writer Identification and Verification using Template Matching Analysis of Texture. *Computer and Information Technology (CIT).* 592 - 597

Akbari, M., Eslami, R., Kashani, M. H. (2012). Offline Persian Writer Identification Based on Wavelet Analysis. *4th International Conference on Bioinformatics and Biomedical Technology, 2012,* 180-186

Aksoy, S., Haralick, R.M. (2001). Feature Normalization and Likelihood-based Similarty Measures for Image Retrieval. *Pattern Recognition Letters 22 (2001)*563-582.

Al-Ma'adeed, S., Elliman, D. and Higgins, C. A. (2002). A Data Base for Arabic Handwritten Text Recognition Research. *Proceedings of the Eighth Internation Workshop on Frontiers in Handwriting Recognition (IWFHR'02), 2002,* 485.

Al-Ohali, Y., Cheriet, M. and Suen, C. (2000). Databases for Recognition of Handwritten Arabic Cheques. *Proceedings of the Seventh International Workshop on Frontier in Handwriting Recognition, September 11-13 2000 Amsterdam,* 601-606.

Bensefia, A., Paquet, T. and Heutte, L. (2005). A Writer Identification and Verification System. *Pattern Recognition Letters, 26(13),* 2080-2092.

Berthold, M. R. (2003). Fuzzy Logic. *In D. J. Hand (Ed.), Intelligent Data Analysis: An Introduction (Second ed.):* Springer.

Bertolini, D., Oliverira, L.S., Justion .E., and Sabourin .R. (2013). Texture –based Descriptors for Writer Identification and Verification. *Expert Systems with Applications.* 40 (6), 2013, 2069–2080.

Bouletreau, V., Vincent, N., Sabourin, R. and Emptoz. (1998). Handwriting and Signature:One or Two Personality Identifier?. *Proc.14th Int'l Conf, Pattern Recognition Brisbane, Australia,* 1,758-1,760.

de Castro, L. N., and Timmis, J. (2002), *Artificial Immune Systems: A New Computational Intelligence Approach,* Springer-Verlag.

de Castro, L. N. and Von Zuben, F. J. Artificial Immune Systems: Part II – A Survey of Applications, *Technical Report – RT DCA 02/00, 2000.* 65.

Der, L. V., Maaten, L. and Postma, E. (2005). Improving Automatic Writer Identification. Proceedings of 17th Belgium-Netherlands *Conference on Artificial Intelligence (BNAIC 2005),* 2005, 260-266.

Djeddi, C., and Souici-Meslati, L. (2011). Artificial Immune Recognition System for Arabic Writer Identification. In *Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on*. IEEE. 159-165

Forrest, S., Perelson, A.S., Allen, L. and Cherukuri, R. (1994). Self-nonself Discrimination in a Computer. *Proceedings of IEEE Symposium on Research in Security and Privacy, 16-18.* 202-212.

Gonzalez, R. and Woods, R. (2008). Digital Image Processing. (3rd edition). New Jersey, USA: Pearson Prentice Hall.

Hu, M. K. (1962). Visual Pattern Recognition by Moment Invariant. *IRE Transaction on Information Theory.* 8(2), 179 - 187.

Jain, A. K., Duin, R. P. and Mao, J. (2000). Statistical Pattern Recognition: A Review Transactions on Pattern Analysis and Machine Intelligence, *IEEE.* 22(1), 4 - 37.

Ji, Z., Dasgupta,D.,(2007). *Revisiting Negative Selection Algorithms, the Massachusetts Institute of Technology,* 15(2), 224-247.

Keijzers, S., Maandag, P., Marchiori, E., & Sprinkhuizen-Kuyper, I. (2013). Image Similarity Search using a Negative Selection Algorithm. In *Advances in Artificial Life, ECAL* (12). 838-845.

Kotoulas, L. and Andreadis, I. (2005). Image Analysis Using Moments. *Proceedings of ICTA'05, 2005 Greece,* 360–364.

Liao, S. X. and Pawlak, M. (1996). On Image Analysis by Moments. *Transactions on Pattern Analysis and Machine Intelligence. 18(3),* 254 – 266.

Muda, A.K., Shamsuddin, S.M., Darus, M. (2006), Bio-Inspired Generalized Global Shape Approach for Writer Identification. *World Academy of Science, Engineering and Technology.*38-44

Mohammed, O.B., Shamsuddin, S.M (2011). Feature Discretization for Individuality Representation in Twins Handwritten Identification, *Journal of Computer Science 7 (7):* 1080-1087.

Mahmoud, A.S., Ahmad, I., Alshayeb, M., Al-Khati, G.W., Parvez, T., Fink, A,G., Märgner, V,. Abid, H (2012). KHATT: Arabic Offline Handwritten Text Databas, *International Conference on Frontiers in Handwriting Recognition,* 449-454

Nemmour, H., Chibani, Y., (2013), Artificial Immune System for Handwritten Arabic Word Recognition, *Innovative Computing Technology (INTECH), 2013 Third International Conference on Digital Object Identifier,* 463-466.

Nguyen, H-T., Vu, N-S., Caplier, A., (2012), How far we can improve micro features based face recognition systems?, *Image Processing Theory, Tools and Applications,* 350-353.

Obied, W., (2010). Handwritten Arabic Writer Identification. *Master, Universiti Teknologi Malaysia, Skudai.*

Parzes, T. A., Mahmoud , S.A. (2012). *Arabic Handwriting Recognition using Structural and Syntactic Pattern Attributes. Pattern Recognition.* 46(1) 141–154.

Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N. and Amiri, H. (2002). IFN/ENIT-database of Handwritten Arabic Words. *In Proceeding of CIFED 2002 Hammamet, Tunisia,* 129-136.

Plamondon, R., Srihari.S.N., (2000). On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *Transactions on Pattern Analysis and Machine Intelligence IEEE, 2002, 22(1),* 63 - 84.

Said, H. E. S., Tan, T. N. and Baker, K. D. (2000). Personal Identification based on Handwriting. *Pattern Recognition.* 33(1), 149-160.

Schomaker, L. (2007). Advances in Writer Identification and Verification. *Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR 2007., 23-26 Sept. 2007 Parana* 1268 – 1273.

Shen, C., Rum, X., Mao,T. (2002). Writer Identification Using Gabor Wavelet. *Proceedings of the 4~ World Congress on Intelligent Control and Automation.* June 10-14, 2002.

Silipo, R. (2003). Neural Networks. In D. J. Hand (Ed.), *Intelligent Data Analysis: An Introduction (Second ed.):* Springer.

Srihari, S. N., Cha, S.-H., Arora, H. and Lee, S. (2001). Individuality of Handwriting: A Validation Study. *Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001, 10 Sep 2001-13 Sep 2001Seattle, WA,* 106 - 109.

Tang, Y., Wu, X., & Bu, W. (2013, June). Offline Text-Independent Writer Identification using Stroke Fragment and Contour Based Features. In *Biometrics (ICB), 2013 International Conference on* IEEE. 1-6.

Wei, Y-G., Zheng, D-l., Wang, y. (2004), Research of a Negative Selection Algorithm and its Application in Anomaly Detection. *Proceedings of the Third International Conference on Machine Learning and Cybemetics,* Shanghai, 26-29.

Zhang, B., & Srihari, S. N. (2003, August). Analysis of handwriting individuality using word features. In *2013 12th International Conference on Document Analysis and Recognition* (2), *IEEE Computer Society.* 1142-1142.