بسم الله الرحمن الرحيم

شبكة المعلومـــــات الجامعية

التوثيق الالكتروني والميكروفيلم

شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الالكتروني والميكروفيلم

﴿ قـــسـم ﴾

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأفلام قد اعدت دون آية تغيرات

يجب أن

تحفظ هذه الأفلام بعيداً عن الغبار

في درجة حرارة من 15 – 20 مئوية ورطوبة نسبية من 20-40 %

To be kept away from dust in dry cool place of
15 – 25c and relative humidity 20-40 %

بعض الوثائق الأصلية تالفة

بالرسالة صفحات

لم ترد بالأصـــل

Ain Shams University
Faculty of Computer and Information Sciences
Department of Scientific Computing

OPTICAL RECOGNITION OF ARABIC CHARACTERS USING NEURAL NETWORK

Thesis Submited For Partial
Fulfillment of Master Degree in
Computer and Information Sciences

By

Ahmed Mahmoud Mahmoud Mohamed
B.Sc. in Computer Science and Statistics

Under the Supervision of:

Prof. Dr. Mohamed. F. Tolba.
Dean of Faculty of Computer and Information Sciences
Ain Shams University

Dr. Gamal. M. Moty.
Department of Scientific Computing
Faculty of Computer and Information Sciences
Ain Shams University

March, 2002

بسم الله الرحمن الرحيم

"و قل ربی زدنی علماً"

صدق الله العظیم

Acknowledgments

During the development of our work we have faced many problems and struggles but we finally managed. First of all, we would like to thank Allah for without HIM nothing in the world is possible

I wish to express my deep gratitude to Prof. Dr. Mohamed F. Tolba for giving me the chance to work under his supervision, for his continuous support and for his periodic and careful evaluation of my work.

I am deeply indebted to Dr. Gamal M. Moty for his valuable guidance, support, help and encouragement during the work presented in this thesis.

I wish to record my thanks to Dr. Sayed Fadel for his help and guidance to give this thesis the suitable final form. I wish also to thank my professors and colleagues for their continuous assistance.

Last but no means least, I am most grateful to my family for their full support and encouragement.

Ain Shams University
Faculty of Computer and Information Sciences
Department of Scientific Computing

OPTICAL RECOGNITION OF ARABIC CHARACTERS USING NEURAL NETWORK

By

Ahmed Mahmoud Mahmoud Mohamed

Abstract

A new approach for Arabic Optical Character Recognition (AOCR) is introduced using sub-word base for segmentation and neural network for feature extractions and classification, and discusses the challenges that Arabic words pose to the implementation of a recognition system.

A hybrid neural network solution, that combines the Principal Component Analysis neural network (PCA) and a Multilayer Perceptrons neural network (MLP), is proposed for recognizing Arabic text. The PCA network is used for feature extraction and MLP network is used for the classification process with specific emphasis on sub-word approach. The hybrid neural network architecture is investigated, and its classification performance is evaluated in the training and testing phases using a database containing 13,800 non-repeated subwords extracted from 1,480,310 Arabic sub-words.

In the present work, we introduce a new sub-word segmentation algorithm that achieves 100% correct segmentation for clean non-Italic text and 98% for Italic text. Then analyzing the collected 13,800 Arabic sub-words, we

reached some results. The first result is concerned with the mean ratio (width: height), which is found to be (2:1). The second result is concerned with the font analysis, which proves that the "Simplified Arabic and the "Time New Roman" can represent other fonts. The third result is concerned with the optimum number of pixels in the normalization standard matrix. Then all these results were integrated together to enhance the normalization process.

Concerning the PCA network, we introduce two dedicated analyses in which we investigate the optimum number of neurons in both the input layer (representing the dimension of the input pattern) and the output layer of the network (representing the number of extracted features) in order to increase the network's qualification.

Another hybrid neural network technique is introduced. It combines both the Self-Organize Feature Map neural network (SOFM), instead of PCA, and the MLP for recognizing the Arabic text. The SOFM neural network has 3 strategic parameters. The first one is concerned with the dimension of the output layer. The second one is concerned with the shape of the shrinking neighbor function. The third one is concerned with the final radius of the shrinking neighbor function. All these parameters should be determined in order to reach the best level of performance. At this point, we introduce results and conclusions through numerous analyses that are documented in detail in this thesis.

Finally, we introduce a comparison between the two hybrid neural networks: PCA/MLP and SOFM/MLP in order to determine the most suitable one for the proposed Arabic Sub-Word Recognition System. The comparison deals with the noisy patterns and target to illustrate the performance of each network at this case. A peak recognition rate of 99% is

achieved at high signal-to-noise ratio (SNR) using the PCA/MLP network, while the rate is dropped gradually to 75% at SNR equal 0 db. In the opposite side, a peak recognition rate of 90% is achieved at high signal-to-noise ratio (SNR) using the SOFM/MLP network, while the rate is dropped gradually to 70% at SNR equal 0 db.

Keywords: Arabic optical character recognition AOCR; principal component analysis PCA; Self-Organizing feature maps SOFM; Multilayer perceptron MLP, Segmentation-Free;

Table of contents

CHAPTER I

INTRODUCTION

CHAPTER I
INTRODUCTION

1.1 Overview

Since the appearance of writing as a means of communication, paper prevailed as the medium of writing. Electronic media has just recently begun to replace paper, because it provides fast and easy access and converses space, in addition to its wide range popularity. The convenience of paper, its widespread use for communication and archiving, and the amount of information already on paper, press for quick and accurate methods to automatically read that information and convert it in electronic form

Character recognition systems can contribute tremendously to the advancement of the automation process and can improve the interaction between man and machine in many applications [1,2], including office automation, check verification, and a large variety of banking [3], business and data entry applications. In addition to other applications including reading postal addresses off envelopes and automatically sorting mail [4], helping the blind to read and reading customer filled forms.

Optical character recognition (OCR) is the branch of pattern recognition that studies automatic reading. The main goal of OCR is to imitate the human ability to read at a much faster rate by associating symbol identities with images of characters. A good typist can type 85 words per minute with an average of 3 mistakes per page. To match those figures, an OCR machine should recognize 99.9% of its input correctly [5] with all errors being rejected and at rate much faster than 5 characters per second. The practical importance of OCR applications, as well as the interesting nature of the OCR