# Analysis and Recognition of Persian and Arabic Handwritten Characters

by

Habib Mir Mohamad Hosseini

B.Sc. in Electronic Engineering, Isfahan University of Technology (1988)
M.Sc. in Digital Electronic, Sharif University of Technology (1991)

Submitted to the Department of Electrical and Electronic Engineering
in fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

UNIVERSITY OF ADELAIDE

November 1997



Signature of Author ..... H. M. Hosseini .....
Department of Electrical and Electronic Engineering
30 November 1997

Certified by .....
Dr. Abdesselam Bouzerdoum
Associate Professor
Thesis Supervisor

# Analysis and Recognition of Persian and Arabic Handwritten Characters

by

Habib Mir Mohamad Hosseini

Submitted to the Department of Electrical and Electronic Engineering
on 30 November 1997, in fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

Though research for designing a machine which can read characters and numerals started more than 90 years ago, problems of recognition of handwritten texts are yet to be completely solved. Even for languages like English or Chinese, for which extensive research has been done, it is probably safe to say that no single scheme is likely to satisfy the requirements in real industrial applications. One of the main reasons is the great variability in handwriting.

The primary goal of this dissertation is to study potential problems of off-line recognition of Persian and Arabic handwritten texts. Specific characteristics of these languages do not allow a direct application of algorithms proposed for the recognition of other character sets. Our study is based on a carefully collected data set containing unconstrained handwritten samples of isolated characters, words, and text from 54 Persian and Arabic speaking writers. Sometimes printed characters and text were used either to analyze the handwriting or to show the difference in recognition of printed and handwritten patterns.

The thesis is divided into three parts. The first part is devoted to analyzing Persian and Arabic handwriting styles. It starts with an introduction to Persian and Arabic writing styles. Then, two of the main problems of a Persian and Arabic handwritten character recognition, namely *similarity* and *variability* of patterns, are addressed. To this end, a geometrical model for distortion analysis of handwritten patterns is introduced, and is then used to investigate the variation of the character patterns. In this model, each distortion source is represented by a transformation matrix operation. Both theoretical and experimental results show that various sources of distortion have different effects on individual characters. Distortion parameters are then estimated for collected handwritten samples of Persian and Arabic characters. This first part is concluded by a comprehensive review on the subject of recognition of printed and handwritten Persian and Arabic texts.

In the second part, we evaluate and test different approaches to feature extraction and classifier design. We also propose some algorithms for feature extraction; in our first approach, we introduced a complex logarithmic transformation technique for invariant feature extraction. This technique is similar to the way the receptors are

distributed in the human retina. This method of feature extraction is then applied to the recognition of both printed and handwritten isolated characters. This feature extraction technique is translation, scale and rotation invariant. For a set of printed Persian and Arabic isolated characters of different scales and rotation ranges, a high recognition rate of 97% was achieved, however, for handwritten characters the system showed a poor performance. The best recognition rates were obtained by using shadow features and a probabilistic classifier, 83% without rejection and 88% with an 11% rejection rate of ambiguous characters.

A new feature extraction technique was developed for recognition of unconstrained handwritten Persian and Arabic numerals. The best recognition rate achieved for a single classifier system was 80%, while using a combined system increased the recognition rate up to 91%. The study of the confusion matrices of the recognition systems revealed that most of the misclassifications were caused by similar digits. The recognition rate was increased up to 94% by rejecting 7% of the patterns.

The elastic matching is among other approaches that have been used to overcome the problem of pattern variation. In a second approach, we used elastic matching technique as a distance measure between the patterns of handwritten digits. Experimental results showed that even these techniques are not capable of completely resolving the problems of ambiguity caused by similar characters and variability of handwriting styles. Some characters become very similar when they are distorted, and hence even elastic matching technique fails to distinguish between these characters. To further improve the performance, context information should be included.

An experiment is done on human recognition of samples of isolated handwritten characters. The best reliability result for the human expert on the collected samples was 0.86. The interesting result is that the best proposed recognition system made almost the same mistakes as human experts; they showed a poor performance in distinguishing between similar patterns. This means that even a human expert is not able to resolve these problems without using context. This led us to the idea of using multiple experts or combination of multiple classifiers techniques to improve the recognition rate of handwritten samples.

In the third part, methodologies for classifier combination are studied. We evaluated three different systems for combining multiple classifiers: weighted voting, linear committee combiner, and a multi-label combiner. In all cases the experimental results showed that the combined system always outperforms all of the individual classifiers. By rejecting ambiguous patterns, both the recognition rate and the reliability improved. Using *a prior* information on the performance of individual classifiers for each class label increased the total recognition rate. The best recognition results achieved by the weighted voting combiner, linear committee combiner, and multi-label combiner were 94%, 96%, and 94% with rejection rates of 28%, 21%, and 24%, respectively.

Thesis Supervisor: Dr. Abdesselam Bouzerdoum
Title: Associate Professor

# Contents


Abstract ii

Acknowledgements vi

List of Figures xi

List of Tables xvi

List of Publications xviii

1 Introduction 1

1.1 Character Recognition Systems . . . . . 2

1.2 Persian and Arabic OCR . . . . . 3

1.3 Definition of the Problem . . . . . 4

1.4 Thesis Outline . . . . . 5

2 Persian and Arabic Handwritings 8

2.1 Introduction . . . . . 8

2.2 Language Description . . . . . 9

2.2.1 Character Sets of Persian and Arabic . . . . . 9

2.2.2 Printed and Handwritten Fonts . . . . . 11

2.3 Problems of Handwriting Recognition . . . . . 15

2.3.1 Segmentation Problem . . . . . 15

2.3.2 Character Primitives . . . . . 19

2.3.3 Number of Classes . . . . . 19

2.3.4 Handwriting Variability . . . . . 19

2.3.5 Confusion of Similar Characters . . . . . 20

2.3.6 Mixture of Fonts . . . . . 21

2.3.7 Problems of Dots and Diacritics . . . . . 21
2.3.8 Lack of Handwritten Data . . . . . 22
2.4 Summary . . . . . 23

**3 Review of the Literature** **24**
3.1 Introduction . . . . . 24
3.2 Research Directions . . . . . 24
3.3 Persian and Arabic Character Recognition . . . . . 26
3.3.1 Psychology of human word recognition . . . . . 27
3.3.2 Data Collection and Analysis . . . . . 27
3.3.3 The Segmentation Problem . . . . . 29
3.3.4 Recognition Systems . . . . . 33
3.4 Summary . . . . . 37

**4 Analysis of Handwriting** **39**
4.1 Introduction . . . . . 39
4.2 Pattern Variability . . . . . 40
4.2.1 Components of Handwriting Style Variability . . . . . 41
4.2.2 Deformation Model . . . . . 42
4.2.3 Variation Analysis of Persian and Arabic Characters . . . . . 49
4.2.4 Estimation of the Parameters . . . . . 56
4.2.5 Deformable Models . . . . . 60
4.3 Pattern Similarity . . . . . 63
4.3.1 Similarity and Confusion . . . . . 64
4.3.2 Similarity and The Deformation Model . . . . . 65
4.4 Approaches to Handwriting Style Variation . . . . . 65
4.5 Conclusions . . . . . 68

**5 Feature Extraction and Character Recognition** **70**
5.1 Introduction . . . . . 70
5.2 Data Acquisition . . . . . 71
5.3 Art of Feature Selection . . . . . 73

7.4.3 Experimental Results . . . . . 131
7.5 Multiple classifiers . . . . . 132
7.5.1 Gating Mixture of the Experts . . . . . 133
7.5.2 Rejecting the Patterns . . . . . 136
7.6 Conclusions . . . . . 137

**8 Conclusions** **139**
8.1 Summary . . . . . 139
8.2 Results and Conclusions . . . . . 140
8.3 Possible Research Directions . . . . . 143

**A Designed Forms** **160**

**B Distortion Characteristics** **164**