FINAL YEAR PROJECT REPORT

URDU OCR

Muhammad Mursaleen Javed	Reg #19187	BSE
Muhammad Aslam	Reg #19185	BSE

Supervised by

Assistant Professor Usman Waheed

Bahria University (Karachi Campus)

2013

Final Year Project Report

URDU OCR

Muhammad Mursaleen Javed	Reg #19187	BSE
Muhammad Aslam	Reg #19185	BSE

Project Advisor Assistant Professor Usman Waheed

Deliverable

Report 1 Volume

Program 1 CD

Bahria University (Karachi Campus)

Acknowledgments

First of all we like to thanks Almighty Allah for giving us such a knowledge and potential to do this paper. This project may not have been completed without the guidance of our kind teachers. We would like to express our special thanks of gratitude to our supervisor Assistant Professor Usman Waheed who gave me the opportunity to do this wonderful project on the topic URDU OCR, which also helped me in doing a lot of Research and we came to know about so many new things I am really thankful to them. Our project examiner Dr. Farah Lakhani for her precious time to guide us for report writing improvements and Prof. Dr. Usman Ahmed (Institute National des Sciences Appliquées de Lyon, France) support toward the project achievement. The prayers of our parents led us to deliver beyond our capabilities. The research media present on internet was also helpful to give directions. Our friends who were always present to motivate us and who helped me a lot in finalizing this project within the limited time frame.

We would like to dedicate this effort to our parents for their love and support.

Abstract

The Idea is to develop an OCR system for Urdu language. There is no such tool available in market to help Urdu publishers to digitalize available Urdu books for Urdu readers. Urdu OCR is a dream goal demands a lot of effort. The available research media was helpful for us to get our directions. We as a team studied available research papers and discussed the possibilities regarding idea implementation. After the discussions we decided to try each and every possibility. We used ACCORD Library for image processing. These possible implementations led us to a decent solution. We focused on pixels and extracted black pixels from the image. Then processed these details for learning purposes (getting new characters / combinations) and then compare these details with the learned data. Provided images are processed to resolve image quality issues, each separate connected ligature is segmented and stored with a unique code. These segmented parts are learned and saved to a defined location and is used to recognize words and symbols. After this process verified words was presented in an Urdu Text editor. In the final quote we would like to get attention and help of the concerned person's regarding future work to make our solution more capable.

.

Table of contents

1. INTRODUCTION 9

1.1 TYPES OF OCR 9
1. Offline handwritten text 9
2. Handwritten text on the Internet 9
3. Machine printed text 9
1.2 USES OF OCR 9
1.2.1 Benefits and Pitfalls of OCR 10
1.2.2 Current state of OCR technology 10

2. BACKGROUND AND LITERATURE REVIEW 12

2.1 URDU WRIING STYLES 12
2.2 URDU FONTS 12
2.3 SEGMENTATION APPROACHES 13
2.3.1 Methodologies 14
2.3.2 Planned / Implemented Methodology 15

3. AIM AND STATEMENT OF PROBLEM 16

4. ANALYSIS AND DESIGN 17

4.1 NAB-OCR: 17
4.1.1 Introduction and history 17
4.1.2 Challenges for Arabic OCR 17
4.1.3 Script Specific Challenges 17
4.1.4 Training procedure: The main goal of the training phase is to prepare the application to be used for text recognition for a certain Arabic script language. Training phase consists of two main steps 18
4.2 METHODOLIGIES 18
4.2.1 Hidden Markov Models 18
4.2.2 Training 18
4.2.3 Recognition 19
4.3 NEURAL NETWORKS 20
4.3.1 Introduction: 20
4.3.2 Learning: 20
4.4 TRANSFER FUNCTION 22
4.4.1 An Example to illustrate the above teaching procedure: 23
4.4.2 The Back-Propagation Algorithm 23
BLOCK DIAGRAM 35

5. IMPLEMENTATION 36

6. TESTING 44

7. RESULTS 45

8. DISCUSSION 46

9. CONCLUSIONS 47
10. FUTURE WORK 48
11. REFERENCES 49