

A Novel Image Matching Approach for Word Spotting

Muhammad Ismail Shah

A Thesis

In

The Department

Of

Computer Science & Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Computer Science at

Concordia University

Montreal, Quebec, Canada

January 2009

© Muhammad Ismail Shah, 2009



Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN: 978-0-494-63295-6
Our file *Notre référence*
ISBN: 978-0-494-63295-6

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.


Canada

ABSTRACT

A Novel Image Matching Approach for Word Spotting

Muhammad Ismail Shah

Word spotting has been adopted and used by various researchers as a complementary technique to Optical Character Recognition for document analysis and retrieval. The various applications of word spotting include document indexing, image retrieval and information filtering. The important factors in word spotting techniques are pre-processing, selection and extraction of proper features and image matching algorithms. The Correlation Similarity Measure (CORR) algorithm is considered to be a faster matching algorithm, originally defined for finding similarities between binary patterns. In the word spotting literature the CORR algorithm has been used successfully to compare the GSC binary features extracted from binary word images, i.e., Gradient, Structural and Concavity (GSC) features. However, the problem with this approach is that binarization of images leads to a loss of very useful information. Furthermore, before extracting GSC binary features the word images must be skew corrected and slant normalized, which is not only difficult but in some cases impossible in Arabic and modified Arabic scripts. We present a new approach in which the Correlation Similarity Measure (CORR) algorithm has been used innovatively to compare Gray-scale word images. In this approach, binarization of images, skew correction and slant normalization of word images are not required at all. The various features, i.e., projection profiles, word profiles and transitional features are extracted from the Gray-scale word images and converted into their binary equivalents, which are compared via CORR algorithm with greater speed and higher accuracy. The experiments have been conducted on Gray-scale

versions of newly created handwritten databases of Pashto and Dari languages, written in modified Arabic scripts. For each of these languages we have used 4599 words relating to 21 different word classes collected from 219 writers. The average precision rates achieved for Pashto and Dari languages were 93.18 % and 93.75 %, respectively. The time taken for matching a pair of images was 1.43 milli-seconds.

In addition, we will present the handwritten databases for two well-known Indo-Iranian languages, i.e., Pashto and Dari languages. These are large databases which contain six types of data, i.e., Dates, Isolated Digits, Numeral Strings, Isolated Characters, Different Words and Special Symbols, written by native speakers of the corresponding languages.

Acknowledgements

First of all, I would like to express my gratitude to my Supervisor, Dr. C. Y. Suen, for his excellent supervision. In addition, I have been his student in two courses, i.e., Pattern Recognition and Expert Systems. I have gained a lot of useful knowledge and ideas from him. Moreover, I am very thankful for the generous financial support he provided during my studies.

The credit of all my success goes to my parents, my brother, my sisters and my wife, who always encouraged me in every difficult situation. It was, in reality, the power of their prayers and honest cooperation, which enabled me to accomplish this challenging task successfully. I will never forget their love and strong support, both moral and financial, which they provided me in a time when they were in a very critical situation, due to the war in our region in Pakistan. I have no words to pay my gratitude to all of them for their generous support.

Furthermore, I am very thankful to Shira Katz, CENPARMI Research Assistant, for her excellent editorial support and to Mr. Nicola Nobile, CENPARMI Research Manager, for his technical support. In addition, I am very grateful to Mr. Wumo Pan, for his sincere cooperation.

Creating a database requires a lot of effort and the collaboration of many people. I would like to pay my gratitude to all the people who helped me in achieving this goal. I am especially thankful to Mr. Fazal Hadi Sarhadi who helped me to collect the handwritten data for Pashto and Dari Databases in Pakistan. I am also thankful to all those writers who filled the data entry forms.

Last but not least, I really appreciate the patience and love of my beloved wife and two small daughters, who are living in Pakistan without me and whom I have always missed very much during the entire course of my study here in Canada.

Dedication

To

My father and mother, i.e., Muhammad Rahim Shah and Iqbal Jehan

My brother and his wife, i.e., Muhammad Ibrahim Shah, and Subhania Shah,

My six sisters, i.e., Nasim Akhtar, Farween Akhtar, Shams Talat, Qamar Talat, Kalsoom
and Chand Talat,

My wife Fouzia Shah and two small daughters, i.e., Hadeeda Shah and Wajeeha Shah

Golden Words

*“Eliminate Belief and Entertain Doubt to
Reach the Apex of Knowledge”*

(Dr. Allama Muhammad Iqbal)

Table of Contents

List of Figures.....	xiii
List of Tables.....	xvii
Abbreviations.....	xix
1. Word Spotting Techniques in Document Analysis and Retrieval –	
A Comprehensive Survey.....	1
1.1. Introduction.....	1
1.2. Background.....	6
1.2.1. Scripts Addressed So Far.....	6
1.2.2. Word Spotting Systems.....	8
1.3. Word Spotting Methodologies.....	10
1.3.1. Pre-Processing.....	10
1.3.2. Features Used.....	13
1.3.3. Image Matching.....	19
1.3.3.1. Similarity/Dissimilarity Measurement.....	19
1.3.3.2. Matching Algorithm/Methods.....	20
1.3.4. Matching Levels.....	28
1.4. System Performance and Evaluation.....	29
1.5. Discussions: Issues and Future Work.....	31
1.6. Conclusion.....	32
2. A New Word Spotting Approach for Image Retrieval from Gray-Scale	
Word Images' Databases.....	34
2.1. Introduction.....	34

2.2. Background.....	36
2.3. Our Approach.....	39
2.3.1. Extraction of Features.....	39
2.3.2. Conversion of Features to Binary Equivalents.....	42
2.3.3. Image Matching.....	44
2.4. Data Set.....	46
2.4.1. Variations in the Data.....	46
2.4.2. Size of Test Data Sets.....	49
2.4.3. Pre-Processing of Data	51
2.5. Experiments.....	52
2.5.1. Overview of Applications.....	52
2.5.2. Analysis of Results.....	54
2.6. Discussions and Future Work.....	60
2.7. Conclusion.....	62
3. The Handwritten Databases of Pashto and Dari Languages.....	63
3.1. Introduction.....	63
3.1.1. Pashto Language.....	64
3.1.2. Dari Language.....	65
3.2. Data Collection.....	66
3.2.1. Data Entry Form.....	66
3.2.2. Writers.....	67
3.3. Data Extraction and Preparation.....	69
3.3.1. Scanning.....	69

3.3.2. Pre-Processing.....	69
3.4. General Overview.....	70
3.4.1. Directory Hierarchy.....	70
3.4.2. Naming Conventions.....	72
3.4.3. Overall Statistics.....	72
3.4.4. Ground Truth Information.....	74
3.5. Data Description.....	75
3.5.1. Letters/Alphabets.....	75
3.5.1.1. Pashto Isolated Letters.....	75
3.5.1.2. Dari Isolated Letters.....	76
3.5.2. Words and Terms.....	77
3.5.4.1. Words in Pashto Database.....	77
3.5.4.2. Words in Dari Database.....	80
3.5.3. Dates.....	82
3.5.4. Special Symbols	83
3.5.5. Numbers/Digits.....	84
3.5.5.1. Pashto Numerals.....	84
3.5.5.2. Dari Numerals.....	85
3.5.6. Numeral Strings.....	86
3.5.6.1. Integer Strings.....	86
3.5.6.2. Real Strings.....	89
3.6. Full Text Handwritten Documents.....	89
3.7. Discussions and Future Work.....	92

3.8. Conclusion.....	93
References.....	94
Related Publications.....	106
Appendix A.....	107
Appendix B.....	109
Appendix C.....	111
Appendix D.....	123

List of Figures

Figure 1.1: Hierarchical representation of various matching methods.....	20
Figure 1.2: Illustration of important regions and lines used for Shape Codes estimation.....	25
Figure 2.1: Dividing the word image into 100 zones, each one of 10 * 10 pixels.....	40
Figure 2.2.a: A Pashto word having Left and Right diagonal lines.....	41
Figure 2.2.b: A Pashto word having several Right diagonal lines.....	41
Figure 2.3.a: Left diagonal area of a zone with radius $R=5$	41
Figure 2.3.b: Right diagonal area of a zone with radius $R=5$	41
Figure 2.4: Representing decimal values of a Gray-Scale Feature Vector by their four binary equivalents.....	42
Figure 2.5: A snapshot of the Database of Features.....	43
Figure 2.6: A sample of the sorted dissimilarities list.....	45
Figure 2.7: Inter-class word variations of 15 different Pashto words.....	47
Figure 2.8: Various classes of words which show more similarities with each other.....	48
Figure 2.9: Intra-class word variations among the various instances of a template.....	49
Figure 2.10: Pre-processed images with no white margins and all scaled to the same sizes.....	51
Figure 2.11: Applications' Overview.....	53
Figure 3.1: Section of Data Entry Form used for recording users' gender and hand-orientation.....	68
Figure 3.2: Hierarchical representation of the Databases' directories.....	71
Figure 3.3.a: A sample of the ground truth information for a numeral string.....	74

Figure 3.3.b: The ground truth information for a handwritten sample of date.....	75
Figure 3.4: Handwritten Samples of 49 characters in Pashto Alphabets.....	76
Figure 3.5: Handwritten Samples of 37 characters in the Dari Alphabets.....	77
Figure 3.6.a: Pashto words used for expressing measurement units and distance terms.....	78
Figure 3.6.b: Pashto words used for expressing measurement units and weight terms..	78
Figure 3.6.c: Pashto words used for expressing measurement units and volume terms..	78
Figure 3.6.d: Pashto words used for writing currency units.....	78
Figure 3.6.e: Pashto words for counting quantities.....	79
Figure 3.6.f: Pashto words used in day-to-day business activities and financial documents.....	79
Figure 3.7.a: Prefixes of measurement units written in Dari scripts.....	80
Figure 3.7.b: Dari words used for measurement units and distance terms.....	80
Figure 3.7.c: Dari words used for measurement units and weight terms.....	80
Figure 3.7.d: Dari words used for measurement units and volume terms.....	81
Figure 3.7.e: Dari words used for writing currency units.....	81
Figure 3.7.f: Dari words used for counting quantities.....	81
Figure 3.7.g: Dari words and terms commonly used in day-to-day activities and financial documents.....	82
Figure 3.8.a: Pashto Dates written in three formats.....	83
Figure 3.8.b: Dari Dates written in three formats.....	83
Figure 3.9.a: Special symbols included in Pashto Database.....	83
Figure 3.9.b: Special symbols included in Dari Database.....	83

Figure 3.10.a: Handwritten Samples of Pashto Numerals.....	84
Figure 3.10.b: Variations found in Pashto Isolated digits database.....	85
Figure 3.11.a: Handwritten Samples of Dari Numerals.....	86
Figure 3.11.b: Variations found in handwritten isolated digits in Dari database.....	86
Figure 3.12.a: Numeral Strings of length 2, selected from Pashto Database.....	87
Figure 3.12.b: Numeral Strings of length 2, selected from Dari Database.....	87
Figure 3.13.a: Numeral Strings of length 3, selected from Pashto Database.....	87
Figure 3.13.b: Numeral Strings of length 3, selected from Dari Database.....	87
Figure 3.14.a: Numeral Strings of length 4, selected from Pashto Database.....	87
Figure 3.14.b: Numeral Strings of length 4, selected from Dari Database.....	88
Figure 3.15.a: Numeral Strings of length 6, selected from Pashto Database.....	88
Figure 3.15.b: Numeral Strings of length 6, selected from Dari Database.....	88
Figure 3.16.a: Numeral Strings of length 7, selected from Pashto Database.....	88
Figure 3.16.b: Numeral Strings of length 7, selected from Dari Database.....	88
Figure 3.17.a: Real Strings and decimal point written in the Pashto numerals.....	89
Figure 3.17.b: Real Strings and decimal point written in the Dari numerals.....	89
Figure 3.18.a: Part of a handwritten document written in Pashto language.....	90
Figure 3.18.b: Part of a handwritten document written in Dari language.....	91
Figure 3.19: The ground truth information for a word within a Dari document.....	91
Figure B.1: The first page of Data Entry Form used for data collection.....	108
Figure B.2: The second page of Data Entry Form used for data collection.....	109
Figure C.1: Sample of handwritten document # 1 written in Pashto scripts.....	110
Figure C.2: Sample of handwritten document # 2 written in Pashto scripts.....	111

Figure C.3: Sample of handwritten document # 3 written in Pashto scripts..... 112

Figure C.4: Sample of handwritten document # 4 written in Pashto scripts..... 113

Figure C.5: Sample of handwritten document # 5 written in Pashto scripts..... 114

Figure C.6: Sample of handwritten document # 6 written in Pashto scripts..... 115

Figure C.7: Sample of handwritten document # 7 written in Pashto scripts..... 116

Figure C.8: Sample of handwritten document # 1 written in Dari scripts..... 117

Figure C.9: Sample of handwritten document # 2 written in Dari scripts..... 118

Figure C.10: Sample of handwritten document # 3 written in Dari scripts..... 119

Figure C.11: Sample of handwritten document # 4 written in Dari scripts..... 120

Figure C.12: Sample of handwritten document # 5 written in Dari scripts..... 121

List of Tables

Table 1.1: Challenges in Typewritten and Handwritten Documents.....	3
Table 1.2.a: Scripts used for word spotting.....	6
Table 1.2.b: List of other languages written in scripts used for word spotting experiments.....	7
Table 1.3: Bounding box estimation at three levels.....	11
Table 1.4: Features, image formats and representations used in word spotting.....	14
Table 1.5: Matching Methods used for Word Spotting.....	23
Table 1.6: Matching levels adopted for word spotting by various researchers.....	28
Table 1.7: Leading Precision Rates achieved in off-line word spotting of various scripts.....	30
Table 2.1: Binary equivalents of the decimal values used in our experiments.....	43
Table 2.2: The handwritten data used in the image retrieval experiments.....	50
Table 2.3: Precision Rates for nine experiments on Pashto and Dari words.....	54
Table 2.4: Individual Precision rates for the nine experiments conducted on Dari words.....	55
Table 2.5: Individual Precision rates for the eight experiments conducted on Pashto words.....	56
Table 2.6: Comparison of results based on various aspects.....	57
Table 3.1: Categorization of Pashto writers based on gender and hand-orientation.....	68
Table 3.2: Categorization of Dari writers based on gender and hand-orientation.....	68
Table 3.3: Overall Statistics of the Pashto Database.....	73
Table 3.4: Overall Statistics of the Dari Database.....	73

Table 3.5: Repetition frequencies of digits in Pashto and Dari Data Entry Forms.....	84
Table A.1: Binary Equivalents of decimal values ranging from 0 to 255.....	106
Table D.1: Detailed Statistics of Dari Database.....	122
Table D.2: Detail Statistics of Pashto Database.....	125

Abbreviations

CBD: City Block Distance

CORR: Correlation (Used to represent Correlation Similarity Measures)

DTW: Dynamic Time Warping

EDM: Euclidean Distance Measure

K-NN: K-Nearest Neighbor

NSHP-HMM: Non-Symmetric Half Plane Hidden Markov Model

P2D-HMM: Pseudo 2 Dimensional Hidden Markov Model

PHMM: Planar Hidden Markov Model (Another name of P2D-HMM)

SLH: Scot and Longuet Higgins algorithm

SSD: Sum of Squared Distances

SC: Shape Context algorithm.

V-Bars: Vertical Bars

Dedication

To

My father and mother, i.e., Muhammad Rahim Shah and Iqbal Jehan

My brother and his wife, i.e., Muhammad Ibrahim Shah, and Subhania Shah,

My six sisters, i.e., Nasim Akhtar, Farween Akhtar, Shams Talat, Qamar Talat, Kalsoom
and Chand Talat,

My wife Fouzia Shah and two small daughters, i.e., Hadeeda Shah and Wajeeha Shah

Golden Words

*“Eliminate Belief and Entertain Doubt to
Reach the Apex of Knowledge”*

(Dr. Allama Muhammad Iqbal)

CHAPTER 1

Word Spotting Techniques in Document Analysis and Retrieval – A Comprehensive Survey

1.1. Introduction

There are millions of very useful books, research papers, theses, journals, and other important historical documents archived in world wide libraries. These documents are highly valuable due to their historical background as well as the information and knowledge they contain. Most of these documents have been degraded to a great extent due to reading and aging processes. Nowadays, there is a great concern about how to preserve and maintain these valuable knowledge assets for future uses. Moreover, it would be desirable to make these sources available for global access over the Internet in the form of future digital libraries.

Usually, within every office, a lot of useful space is occupied by a huge amount of record books and other documents which can be used for other purposes. In addition, the administrative records in the government offices, universities, colleges, hospitals and other social and business organizations must be preserved and maintained. Most of these documents/files contain very important information such as customer and employee profiles, expense and revenue details, etc., accumulated over many years. The preservation, management and maintenance of these administrative documents is a great challenge. The demand for an efficient online system to access these important records and valuable sources of knowledge increases every day. With recent developments in information technology, it has become easily possible to scan and store these valuable

documents. However, the actual problem is the retrieval of these documents, that is, after digitization, how to reliably and efficiently access these documents for some specific information. There is a need for a robust system which can effectively manipulate and classify the contents of these documents.

For the last several decades the Information Retrieval community has been trying to find a well-organized way for indexing these huge amounts of documents and for retrieval of only certain relevant documents based on each user's query [57, 58, 65]. Optical Character Recognition (OCR) techniques have been applied by some researchers [39, 68, 69] to retrieve specific documents from document databases. However, these techniques are not effective because most of the documents within libraries and offices are either in a much degraded condition or written in languages for which presently no OCR technologies exist.

Traditional OCR technology strictly depends on careful analysis of page layouts, segmentation of words into isolated characters and conversion of those characters into machine-readable texts followed by a recognition process. OCR techniques have been applied to limited areas, e.g., bank cheque recognition [70, 71], and postal code/address recognition [72, 73, 74]. These applications require complex classifiers trained with a small size of domain-specific lexicons. For good quality typewritten documents, the existing OCR technologies can perform satisfactorily. However, these traditional OCR techniques can not be used to retrieve information from degraded documents with unlimited/unseen vocabulary and complex layout of pages. As shown in Table 1.1, there are many challenges, not only in handwritten but also in machine printed documents.

These problems hinder the use of OCR techniques for the purpose of creating digital libraries for online access and reliable retrieval of client-specific information.

Table 1.1: Challenges in Typewritten and Handwritten Documents.

Challenges in Typewritten / Printed Documents	Challenges in Handwritten Documents
Variable Font Sizes and Styles such as <i>Italic</i> , Bold , and <u>Underlines</u> . Different types of Font Faces such as Tahoma, Broadway, and Courier. Use of Charts, Diagrams, Tables etc.	Non-uniform inter/intra word spacing, Slanted characters, Skewed lines, Non-uniform inter-line spaces, Uneven ink distribution, Overlapping and touching characters, Multiple writers, Stylistic variations, Corrections, Cutting lines, and ink bleeds etc.

Moreover, for oriental languages such as Indo-Iranian languages written in cursive scripts, the traditional OCR technology can not perform well, even in the printed documents. For these scripts, it is very difficult and in some cases impossible to correctly segment words into isolated characters. Books and documents archived at libraries and offices are usually written in many languages and may contain complex diagrams, charts, tables, equations, symbols and images. Due to this fact, these documents may not be easily converted into machine-readable form required for OCR techniques.

A more robust and effective approach is required for managing these multilingual typewritten and handwritten documents, i.e., indexing and retrieving. Word spotting, originally adopted by the speech recognition community [55, 56], has been practiced as an alternative technique to OCR technology by many researchers for the last one and half a decades. This approach has been used for experiments on indexing handwritten documents [1, 13, 19, 29] as well as for information retrieval from printed [12, 21, 43, 51] and handwritten [8, 9, 41, 46] document image databases. Word spotting techniques have been shown to be more effective than OCR, especially for degraded and handwritten

documents. Decurtins [16] has compared a keyword spotting system with OCR and has shown that word spotting system is more robust. Yue, et al. [43] have also compared their word spotting method with an OCR tool and have reported that their word spotting system is more effective and efficient.

The Word Spotting approach, also called “word searching”, is used to detect and locate instances of the given template/query image in the document image databases. When documents in which the instances of the template are detected, they are retrieved and presented to the user. The basic idea of word spotting is that a template image is selected from a set of predefined keywords, i.e., words of interest, and then a search is initiated to find out its other instances in the target set of digitized documents. The process is totally based on image matching and conversion of whole documents to machine readable texts, i.e., ASCII codes, and machine recognition are not involved at all. This factor makes the approach more flexible and suitable for indexing and retrieval of degraded and historical documents written in multiple languages.

The effectiveness and robustness of any Word Spotting system largely depends on pre-processing methods, types of features extracted and the matching methods/algorithms. The Pre-processing may include noise removal, normalization, skew detection, identification and location of bounding boxes of target images and feature extraction. Various methods can be used to identify and locate bounding boxes of the target images, e.g., Morphological Operations, Scale Space Techniques [63, 64] and Gap-metrics [60]. The use of various types of statistical and structural features based on image representations, that is, skeletons, contours and pixels, and image formats (i.e., binary or grey-level) have been reported by the research community. Dynamic programming

techniques have been widely used for matching visual similarities between the template and target images. These techniques include Viterbi decoding on HMM [12, 20, 21, 28], Dynamic Time Warping [1, 15, 29, 32, 51], and Inexact Feature Matching [43, 45], etc. In addition traditional distance measurement techniques have also been applied, e.g., City Block [2, 50] and Euclidean Distance Measurement [14, 34, 42, 38].

The purpose of this survey is to highlight the importance of word spotting by briefly reviewing the various efforts made in the past. We have studied the word spotting techniques used in both printed as well as handwritten documents. The readers may find similarities and differences among various approaches and thus could pinpoint the effective methods that could be used in similar domains. One could find out the possibilities of applying these methods for other languages which have not yet been addressed by the research communities.

The rest of the chapter has been organized as follows. Section 1.2 gives a brief background review, i.e., various scripts addressed by the word spotting community and some Word Spotting Systems. Section 1.3 is the major section of this chapter, which describes the basic components of the word spotting techniques. Its four subsections include the various word spotting approaches that have been used for pre-processing, features extraction, image matching and results evaluation, respectively. In Section 1.4, we describe system evaluation techniques in word spotting, i.e., precision and accuracy. In Section 1.5, we discuss some basic word spotting issues and the future possible work in this regard. Section 1.6 provides conclusive comments based on the entire study presented in this survey. We hope that this work will prove to be a rich source of useful information for future researchers interested in word spotting.

1.2. Background

1.2.1. Scripts Addressed So Far

In this section, we present an overview of various scripts that have been addressed by the research community in the domain of word spotting. So far, word spotting techniques have been applied to various scripts, i.e., Latin, Arabic, Greek, Chinese, Devanagari, and Amharic, as shown in Table 1.2.a.

Table 1.2.a: The scripts used for word spotting.

Publications	Language	Scripts	Domain	
			Type	Print
Rath et al.[1, 13], Manmatha et al.[19], Cao et al.[3], Rothfeder et al.[35], Tomai et al.[39], Adamek et al.[4], Srihari et al.[46], Zhang et al. [41]	English	Latin	Off-line	Handwritten
Keaton et al.[18], Kotcz et al.[15]	Spanish			
Leydier et al.[6, 10], Christophe et al. [53]	French			
Ntzios et al. [5]	Greek			
Srihari et al.[8, 9, 46], Al-Khatib et al.[34], Shahab et al. [38], Saabni et al. [85]	Arabic	Arabic		
Jain et al.[33], Burl et al.[17]	English	Latin	On-line	
Chen et al.[12, 21, 28], Kuo et al.[20, 24], Linlin et al.[7], Spitz [26], Jeffrey et al.[25], Jawahar et al. [51], Balasubramanian et al.[52], Shijian et al.[37], Yue et al. [43, 45]	English	Latin	Offline	Type-written
Josep et al.[54]	Spanish			
Yue et al. [14, 42]	Chinese	Chinese		
Soo et al. [49], Park et al.[50], Cho et al. [48]	Korean			
Gatos et al.[11], Konidaris et al.[2]	Greek	Greek		
Srihari et al.[46]	Sanskrit	Devanagari		
Jawahar et al. [51], Balasubramanian et al.[52]	Hindi			
Jawahar et al. [51]	Amharic	Amharic		
Saabni et al. [85]	Arabic	Arabic		

Most of these scripts differ from each other, based on various factors such as: number of characters, writing direction and cursiveness. For example, languages in Latin scripts are written from left to right in horizontal direction only, either cursively or non-cursively. However, languages in Arabic Scripts are written from right to left, in a

horizontal direction, and are totally cursive. On the other hand the Chinese scripts, e.g., Chinese and Korean, have thousands of characters and are written in two dimensions, i.e., left-to-right horizontally and top-to-bottom vertically. Languages written in Devanagari scripts are written horizontally, from left-to-right, in a complex cursive way. Greek and Amharic scripts are written from left-to-right and are non-cursive. Furthermore, each isolated character of the Chinese scripts has certain meanings/semantics whereas the isolated characters of other scripts, e.g., Latin, Arabic, Greek and Devanagari, etc. have no meanings.

Table 1.2.b: List of other languages written in scripts used for word spotting experiments.

Scripts	Letters	Well-known Languages written in the corresponding scripts	Languages used for Word spotting
Arabic	28	Arabic language, Balochi, Kashmiri, Kazakh, Kurdish, Malay, Pashto, Persian (Farsi/Dari), Punjabi, Sindi, Urdu	Arabic
Latin	26	Albanian, Danish, English, Filipino, Finnish, French, German, Hungarian, Indonese, Italian, Kurdish, Portuguese, Rumanian, Somali, Spanish, Swedish, Turkish, Uzbek, Vietnamese	English, French, Spanish
Greek	24	Greek, Iberian	Greek
Devanagari	52	Bhojpuri, Gondi, Limbu, Marathi, Mundari, Nepali, Newari, Rajasthani, Sanskrit, Santali, Hindi*	Hindi, Sanskrit
Chinese	3000 +	Japanese, Korean, Chinese	Korean, Chinese
Amharic	231	Amharic language, Geez, Oromo, Tigre, Tigrinya	Amharic

As shown in Table 1.2.b, there are many other languages written in scripts for which word spotting techniques have been successfully applied. The proposed techniques for word spotting in one language can be easily applied to other languages written in the corresponding scripts. Most of the researchers have focused on only one type of script, e.g., Latin [1, 13, 19, 35], and they have not tested their methods on other scripts. However, some authors have tried to test their methods on more than one type of scripts.

* Urdu and Hindi are considered to be the two names for the same language. Hindi is used in India and is written in Devanagari scripts, whereas the name Urdu is used in Pakistan and is written in Arabic scripts.

For example, Leydier et al. [6, 10] have applied their proposed method to two different types of scripts i.e., Latin and Arabic. Srihari et al. [46] have applied their method to three types of scripts, i.e., Arabic, Latin and Devanagari. In addition, Jawahar et al. [51] have also used the same method for three different scripts, i.e., Latin, Devanagari and Amharic.

1.2.2. Word Spotting Systems

So far, various software systems used for word spotting have been reported by the document retrieval research community. Here, we give a brief summary of some word spotting systems:

Srihari et al. [8, 9] presented an interactive software system called CEDARABIC designed for word spotting in handwritten Arabic documents. The proposed system provides a Graphical User Interface for inputting a query and displaying the results of the query. The working mechanism is like a dictionary look-up in which the user enters a query word, typed in English text, and the system finds its Arabic equivalent words in the database. The database contains pre-segmented handwritten Arabic documents accompanied with well-organized ground truth information such as the alphabet sequence, meaning and pronunciation of each word. In the first step, the word image is retrieved from the reference data set and used as a template. In the second step, the documents which contain instances of the given template are retrieved and ranked. The ranking is based on the value of similarity between the template and the instances found in each corresponding document.

Srihari et al. [46] made some modifications to the system presented by Srihari et al. [8, 9]. They extended the system's functionalities to perform word spotting in three

languages, i.e., handwritten Arabic and English as well as typewritten Sanskrit (Devanagari) documents. Furthermore, the first version of the system [8, 9] allows the users to enter the query only as an English text whereas in the modified version [46], the users can also enter word images as queries.

Al-Khatib et al. [34] presented an interactive word spotting system for indexing and retrieving historical Arabic manuscripts. The system is based on the framework of a digital library described by Shahab et al. [38]. It has several important functionalities, such as pre-processing, feature extraction, image matching and a well-designed Graphical User Interface for input and output applications. The system uses the user's feedback to index the manuscript pages, which enables efficient retrieval of the relevant documents in future queries.

DeCurtins et al. [23] presented a software system, called SCRIBBLE, for word spotting in machine printed English documents. The system requires some sort of pre-training in the form of an *a priori model*, which specifies the usage of patterns for various words. The word shape codes along with a voting technique are used to compute the likelihood of a given word within a text line. This method is robust due to its voting scheme and can be applied to noisy documents, where the possibility of missing or fake features is higher. DeCurtins [23] compared SCRIBBLE with an OCR system and reported that this word spotting system has a higher performance.

Jawahar et al. [51] proposed a prototype of a word spotting system to retrieve relevant documents from a large collection of printed documents. The documents were initially indexed by clustering the similar words, whereby each cluster was assigned a weight and relevance to the corresponding documents. The system has been applied to

documents printed in three types of scripts, i.e., Latin, Devanagari, and Amharic. The system works at two levels: first, feature vectors are aligned using dynamic programming and then structural features are used to compare the shapes of the images. The same framework has been implemented by Balasubramanian et al. [52], however, they applied it to documents printed in Latin and Devanagari scripts only.

Rath et al. [45] presented a search engine for retrieving handwritten historical manuscripts written in English, i.e., George Washington's manuscripts. The system was initially trained with an annotated set of documents and was used to retrieve documents from the target set of documents based on users' queries. The target set contained untranscribed pages written by the same writer.

1.3. Word Spotting Methodologies

1.3.1. Pre-processing

The robustness of any word spotting system is directly related to and dependent upon the pre-processing phase. This phase usually involves noise removal, skew correction, bounding box estimation and feature extraction. Based on these pre-processing methods, the word spotting techniques can be broadly divided into two groups. In the first group, the entire document is considered as a single image and no page segmentation is required [6, 10]. The given document images are scanned in the direction of writing to find out the instances of a given template based on shape similarities. In the second group, the document image is considered as a collection of sub-images, i.e., words and/or characters. Before keyword searching is initiated, the given document image is segmented into its sub-images.

Approaches for estimation of bounding boxes within a document image can be either Top-down or Bottom-up [42]. In the Top-down methods, the layout of the documents is carefully analyzed and used to segment the document images into various sections based on the type of contents it contains such as text, graphics, tables, charts, etc. The search for a given template image is initiated in the text portion which is usually divided into headlines and body text. This approach is effective in the domain of printed documents with a constant specific format [42]. In the Bottom-up approaches, the connected components within a given document are initially detected and analyzed. The documents are segmented by estimating the bounding boxes of the connected components. Estimation of the bounding boxes within a text document image is a very important step as the overall accuracy and throughput of the system depends on it. For printed documents with stable and uniform layouts, bounding box estimation is not a big problem, however, for handwritten documents it is a real challenge. In the word spotting literature, the bounding boxes have been extracted at three levels, as shown in Table 1.3, i.e., Lines, Words or Characters.

Table 1.3: Bounding box estimation at three levels.

Publications	Bounding Box Estimation
Rath et al. [1, 13, 29], Rothfeder et al. [35], Adamek et al.[4], Manmatha et al. [19, 36], Al-Khatib et al.[34], Shahab et al. [38], Chen et al.[21, 28], Tomai et al. [39], Jain et al.[33], Jawahar et al. [51], Balasubramanian et al.[52], Srihari et al.[8, 9, 46]	Word Level
Cao et al.[3], Yue et al. [14, 42], Park et al.[50], Soo et al. [49]	Character Level
Kotcz et al.[15], Chen et al.[12]	Line Level

In the line oriented approaches, the page is first segmented into lines. Each line is then scanned along the direction of writing to spot a word similar in shape to the given

word model/template [15]. The widely used approach is based on estimation of word level bounding boxes. The document images are segmented into words using connected component analysis. The template/keyword is either compared with every word image in the document image [33] or with only those images which have aspect ratios equal/close to the keyword [1]. In the latter case, a method called pruning is used according to which unlikely instances of the keywords are not considered for matching, i.e., words having aspect ratios which are either very large or very small as compared to the given keywords [1]. In the character level approaches, the bounding boxes are estimated for each character. This approach has been widely used for word spotting in printed Chinese [14, 42] as well as Korean [48, 49, 50] documents. Furthermore, this approach has also been employed in the word spotting approaches where shape codes are derived by using the character level shape information [26].

In addition, Keaton et al. [18] have used a method in which a set of candidate locations are identified and extracted, i.e., text blocks that may contain the instances of the keywords. Correlation strength is used to rank different candidate locations and those with high correlations are selected for pre-processing. The instances of keywords are then searched by scanning these locations.

Page segmentation for word spotting is a rich area of research and so far, many useful approaches have been reported. For example, Manmatha et al. [63, 64] have proposed a scale space technique for automatically segmenting historical handwritten documents. Mahadevan et al. [60] have used Gap metrics for segmenting words within handwritten text lines. Laurence et al. [59] have surveyed various segmentation

approaches suitable for segmentation of historical documents. In this paper, we are not explaining the segmentation methods as it is a vast field of research in its own.

1.3.2. Features Used

In this section, we present an overview of the various important features used by different researchers for word spotting. In addition to feature selection, other important factors such as image format and representation must be carefully considered, as they have a great impact on the system's overall performance. Furthermore, the extraction of features also depends upon the image format and image representation used. As shown in Table 1.4, most of the researchers have used the features extracted from Binary Images. However, most of the word spotting work done in the domain of historical handwritten documents is based on Gray-Scale Images [1, 6, 10, 35]. The reason for this is that binarization of word images results in the loss of important information which can lead to unsatisfactory results.

There are three types of well-known image representations, i.e., Pixels, Skeletons and Contours. From Table 1.4, we can find that about 85% of researchers have adopted Pixel representation for word spotting. Similar to image format, image representation can also affect the system's performance in terms of computational cost and information losses. For example, using Skeletons can significantly reduce the computational cost but it can cause a loss of useful information during binarization and thinning operations. In Pixel representation, extraction of most of the features depends upon correct estimation of each word's baseline. On the other hand, baseline estimation is not required in Contour representation. The selection of representation largely depends upon the features used in the process. For example, for extraction of point features, i.e., end points, intersections,

etc., and cavity* features the images are required to be skeletonized. Similarly, for extraction of projection profiles the Pixel representation is necessary.

Table 1.4: Features, image formats and representations used in word spotting.

Publications	Features Used	Image Format	Representation
Rath et al.[1]	Projection profiles, Upper/lower word profiles, Background-to-ink transitions	Gray Scale	Pixels
Rothfeder et al.[35]	Corner features		
Cao et al. [3]	Gabor features		
Leydier et al. [6, 10]	Gray level, Gradient orientation		
Chen et al. [12]	Column pixel values		
Kotcz et al.[15]	Upper/lower word profiles, Background-to-ink transitions	Binary	
Gatos et al.[11], Konidakis et al.[2]	Mesh features, Upper/lower word profiles		
Srihari et al. [8, 9, 46], Zhang et al. [41]	Gradient-based binary features (GSC Binary Features)		
Al-Khatib et al. [34], Shahab et al. [38]	Angular line, Concentric circle, 15-point DFT transform of vertical/horizontal profiles, Height, Width, Area and Aspect Ratios		
Jawahar et al. [51], Balasubramanian et al.[52]	Projection profiles, Upper/lower word profiles, Background-to-ink transitions, Moments, Mean, Black pixel distribution, Strokes' curvature		
Kuo et al. [20, 24]	Pixel's value, Transactional information, Pixel's relevant position in row and column		
Yue et al. [43, 45]	Straight stroke lines, Traversal features		
Yue et al. [14, 42]	Strokes' density		
Saabni et al. [85]	Geometric Features		
Soo et al. [49]	Mesh features		
Spitz [26]	Ascenders and Descenders, East concavity feature		
Jeffrey et al. [25]	Spatial Moments: means, variances		
Shijian et al. [37]	Holes, water reservoir, ascenders/descenders		
Linlin et al. [7]	Straight vertical/non-vertical strokes, Ascenders and Descenders		
Park et al.[50]	Mesh features, Upper/lower and right/left word profiles, Wavelet coefficients		
Christophe, et al. [53]	Image column height, Neighbor hood size		
Keaton et al.[18]	Projection profiles, Cavity features	Binary	Skeletons
Josep et al. [54]	Skeleton points		
Ntzios et al. [5]	Open/closed concavity features		
Adamek et al.[4]	Multi-scale convexities/concavities features	Binary	Contours
Chen et al. [21, 28]	Upper/lower contours, Auto-correlations	Gray Scale	

* For extraction of cavity features, skeletonization is effective but not strictly required.

Similar to character recognition, in word spotting the selection of proper and effective features plays a vital role, especially in historical and degraded handwritten documents which contain various types of noises. Rath et al. [13] have explained some of the features used for word spotting in historical manuscripts. In this paper we have given a brief overview of some of the features used for word spotting:

Projection Profiles [13]: These features are calculated by summing the number of black pixels in each image column. In this way, the length of the feature vector is equal to the total number of columns, where each vector value represents a single column. The black pixels can either be counted for the entire image column or part of it, i.e., partial projection profile. For extracted partial projection profiles, three zones of the image are considered as shown in Figure 1.2, in Section 1.3.3.2, i.e., ascenders-zone, x-zone, and descenders zone. The projection profiles are calculated for each zone, which is called Upper projection profile, Middle projection profile and Lower projection profile, respectively.

Word Profile [13]: Word profile features can be calculated from four directions of a given word image, i.e., upper, lower, right and left. The first two are calculated by scanning the image column-wise whereas the latter two are calculated by scanning the image row-wise. The distance of the first ink pixel, within each image column, from the upper word boundary is called the upper word profile, whereas that from the lower word boundary is called the lower word profile. Similarly, the distance of the first ink pixel, within each image row, from the left boundary is called the left word profile while that from the right is called the right word profile. If a column or row has no ink pixel at all, then linear interpolation is used to calculate the feature value.

Mesh Features [49]: The image is logically divided into a fixed number of zones, and for each zone the black pixel density is calculated, i.e., total number of ink-pixels. The length of the feature vector is equal to the total number of zones.

Cavity Features [18]: These are features which represent the gaps between strokes of word. These features capture local variations and are used to distinguish different words with similar general shapes. A region point which is bounded by the character strokes on at least three directions/sides is called a cavity. Keaton et al. [18] have used six types of cavities, i.e., east cavity, west cavity, north cavity, south cavity, center cavity and hole. The first four are named after the side on which the region is not bounded. A completely bounded region is called a hole, while a region which is bounded on four sides (but is not a hole) is called center cavity.

In the word spotting literatures the cavities have been used in different ways. For example, Keaton et al. [18] have considered cavities as principal features for image matching. Whereas, Ntzios et al. [5] do not use cavities as principal features but instead extract various features based on open and closed cavities and use them for classification of characters and ligatures. These features consist of length of various protrusive segments, slopes and opening angles of cavities.

Gradient Orientation [6, 10]: A feature which describes the shapes of words by depicting the strokes' true local structure and the orientation of characters' contours. Usually, it is a feature vector, computed from the gradient of grey level, which points in the direction in which the grey level's value increases to maximum and whose value represents the rate of change.

GSC Features [61]: These are the combination of three features, i.e., Gradient, Structure and Concavity features. These features measure the characteristics of an image at local, intermediate and global ranges respectively. The Gradient features measure edge curvature in the neighborhood of a pixel and provide useful information about the stroke shape. They are then further extended to a longer distance by Structural features in order to achieve important information about stroke trajectories. The stroke relationships at a global scale are detected by Concavity features, i.e., across the entire image. For details of how these features are calculated see Favata et al. [61].

Angular Lines Features [38]: An x-y plane is created, which has its origin at the centroid of the word image. Each quadrant of this plane is divided into two equal sections (45 degrees each), thus forming 8 equal regions. In each region, the number of pixels is counted which results in the creation of an 8-valued feature vector. The number of pixels in each region is divided by the total number of pixels in the word image for the purpose of normalization.

Concentric Circle Features [38]: Several concentric circles are drawn by considering the centroid of the word image as the center. The number of pixels between the two consecutive circles are counted and used as feature values. In [38], four circles have been drawn, however, this number can vary depending upon the preference of the user. For computing the number of pixels between two circles, the total number of pixels of the inner circle is subtracted from the total number of pixels enclosed by the outer circle.

Upper/Lower Contour [21, 28]: The upper contour is the distance from the upper outline of a word to the top of the bounding box. The distance from the lower outline of a word to the bottom of the bounding box is called its lower contour. If characters within the

word are not touching each other, then the lower contour follows (i.e., touches) the top of the bounding box and the upper contour follows (i.e., touches) the bottom of the bounding box.

Ascenders/Descenders: These features depend upon the baseline and the x-line of the word image, as shown in Figure 2.2. The ascenders are the characters extended above the x-line while descenders are the characters extended below the baseline. These features have been largely used for word spotting in Latin scripts. Like cavity features, these features are also used in several ways, e.g., they can either be used as principal features [21, 28] or initially used to define a set of shape codes [7, 26, 43, 45]. In the latter case, the matching is based on the derived shape codes and not on the features directly, as will be discussed in Section 1.3.3.1.

In addition to the selection of proper features, the pattern of their usage is also important for getting satisfactory results. Using various features in a combined way can give a high performance as compared to the corresponding features individually/separately. Rath et al. [1] and Keaton et al. [18] have shown that using a combination of features can produce a high performance and more satisfactory results. In addition, the best choice of features to be used also depends on the matching method. For example, the profile features have been proved to be more effective when using the Dynamic Time Warping method because these features can be easily represented as a time series.

1.3.3. Image Matching

Image matching is not only the most important, but also the most difficult phase in word spotting. In the following sub-sections, we describe image matching concepts and various approaches adopted by researchers for word spotting.

1.3.3.1. Similarity/Dissimilarity Measurement

Two images are usually matched based on either similarity or dissimilarity functions. An image I_{target} is considered to be an instance of a given image $I_{template}$ if the output of their similarity function is greater than some predefined threshold, i.e.,

$$[I_{template} \equiv I_{target}] \rightarrow [Similarity(I_{template}, I_{target}) > Threshold]$$

On the other hand, an image I_{target} is considered to be an instance of a given image $I_{template}$ if the output of their distance/dissimilarity function is lower than some predefined threshold, i.e.,

$$[I_{template} \equiv I_{target}] \rightarrow [Dissimilarity(I_{template}, I_{target}) < Threshold]$$

Some of the well-known similarity measurement algorithms are: Dynamic Time Warping (DTW) [1], Shape Context (SC), and Correlation Similarity Measure (CORR) [31]. In addition to CORR, Tubbs [31] has also defined seven other types of similarity measures for matching two binary patterns.

The well-known Minkowski distance [67], i.e., City Block Distance (1 -norm) and Euclidean Distance (2 -norm), have been widely used by researchers, as shown in Table 2.6, for matching dissimilarity between two images. City Block Distance (CBD) and Euclidean Distance Measure (EDM) are very useful distance measurement techniques. These techniques are not only used as stand-alone matching methods but other image matching methods, such as DTW and K-NN, also depend on some distance function

defined by either City Block Distance (CBD) [15, 18] or Euclidean Distance Measure (EDM) [1, 33, 51, 52].

1.3.3.2. Matching Algorithms/Methods

In the word spotting literature, there are various types of algorithms and methods used for image matching, as shown in Table 1.5. These matching methods can be represented hierarchically, as depicted in Figure 1.1.

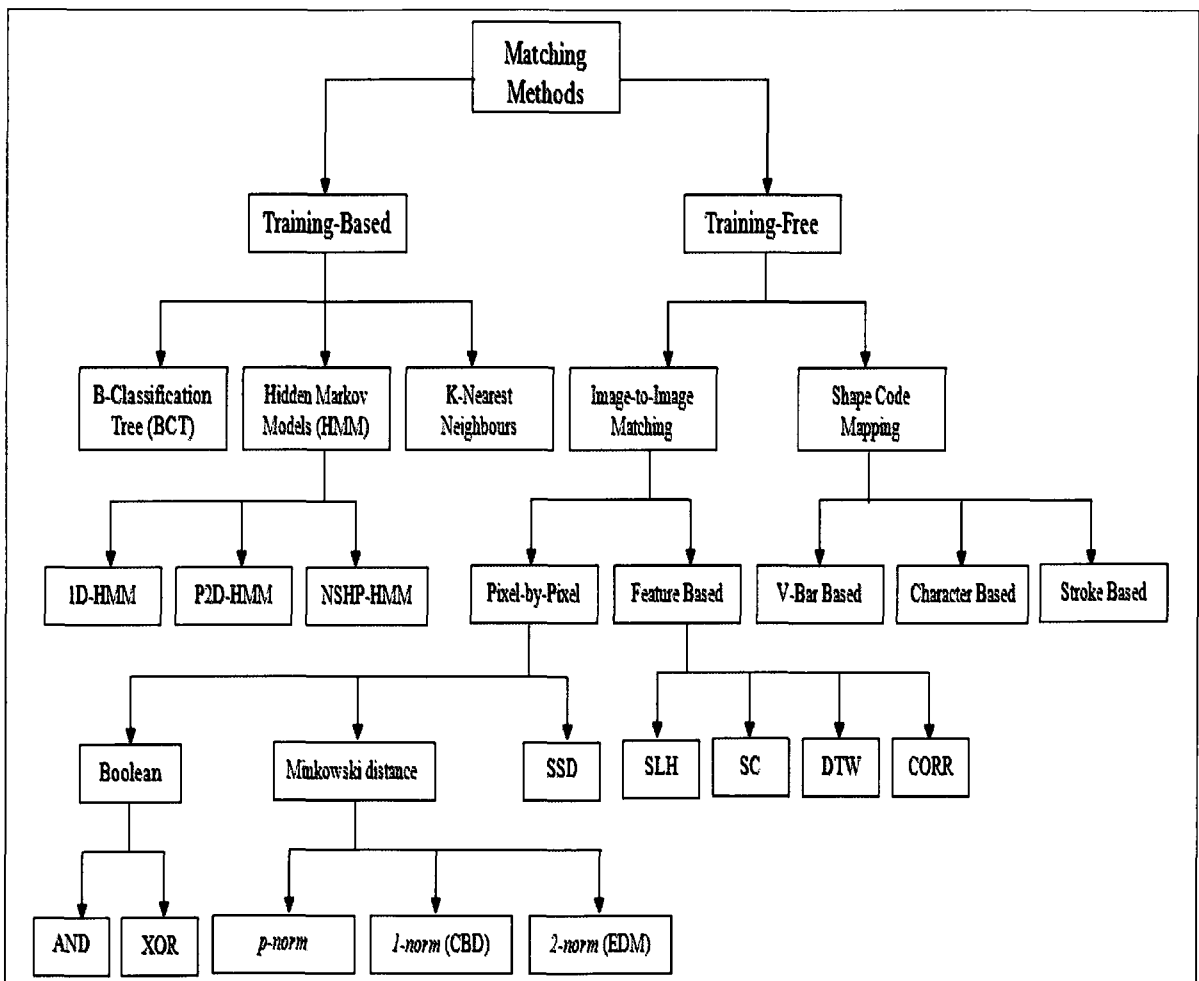


Figure 1.1: Hierarchical representation of various matching methods.

We have categorized the various image matching techniques into three groups. The first two groups represent those matching methods which are training-free, i.e., Image-to-

Image Matching and Shape Code Mapping. The third group represents those matching methods which require a so-called training step, i.e., Hidden Markov Model (HMM), K-nearest Neighbor (K-NN). Table 1.5 provides a summary of various publications where these types of matching methods have been used for word spotting purposes. Here we give a brief summary:

Group 1: Image-to-Image Matching: This type of matching is also called template matching, where a pre-defined template/keyword is matched with the target images within a given set of documents. The matching can be either pixel-by-pixel or feature-based, where initially some features are extracted from the two images and then aligned and compared to find similarities between the template and the target image. The image may be a word, phrase or character depending on the required applications.

In the pixel-by-pixel matching methods, the template and target images are raster scanned and the distance between the corresponding pixel values is calculated. The Minkowski distance has been widely used for this type of matching, i.e., City Block Distance [49, 50, 11, 2] and Euclidean Distance [14, 42, 34, 38]. In addition to the Minkowski distance measurement techniques, some other pair-wise image matching methods have been reported. For example, Manmatha et al. [19] have used the XOR method for matching two images, where the two binary images are aligned and a difference image, i.e., XOR-image, is calculated by XORing the pixels of the two images. The matching cost is represented by the number of difference pixels. The XOR method is similar to Euclidean Distance Measure (EDM), however, in EDM the difference pixels in a larger cluster/group result from structural differences between the two matching images, hence their weight is heavier. In the Sum of Squared Differences (SSD) method,

the two images are translated relative to each other and the matching cost, i.e., minimum cost, is calculated based on the Sum of Squared Differences [32].

In the feature-based image-to-image matching, a fixed number of features are extracted and represented as feature vectors. Similarity between the template and target images is measured by comparing their corresponding feature vectors. This approach has been adopted by many researchers for word spotting and has proven to be more successful and robust than pixel-by-pixel image matching techniques. Rath et al. [1] and Rothfeder et al. [35] have conducted comparative experiments and have reported that the feature-based techniques are more robust and tolerant to handwriting variations than the pixel-by-pixel matching methods.

Scott et al. [62] have presented an algorithm which matches features extracted from two patterns. This algorithm has been named after its developers, i.e., Scott and Longuet-Higgins as the SLH algorithm. The SLH algorithm has been used by various researchers [19, 32] for word spotting in historical documents. In this method, sample points are taken from the template and target images and an affine warping transform is recovered between them. The matching cost is represented by the residual between the template points and the warped candidate points [32].

Belongie et al. [30] introduced a shape similarity measurement algorithm called Shape Context (SC), which is based on taking sample points from the outlines of the two images which are to be matched. Corresponding sample points in the two images are recovered by assigning a shape context histogram to each corresponding point. These correspondences describe the distribution of sample points in the shape with respect to the reference sample point, i.e., the point at which it is generated. The candidate image is

warped by iteratively performing the matching operation and the cost associated with the chosen correspondences is used to determine the matching cost [1, 32].

Table 1.5: Matching Methods used for Word Spotting.

Publications	Major Features Used	Matching Methods/Algorithms
Rath et al.[1], Jawahar et al. [51], Balasubramanian et al. [52]	Projection profiles, upper/lower word profiles, background-to-ink transitions	EDM, DTW
Jain et al.[33]	Height, direction and curvature of strokes	
Soo et al. [49]	Mesh features	City Block Distance (CBD)
Park et al.[50]	Mesh features, upper/lower and left/right word profiles, wavelet coefficients	
Gatos et al.[11], Konidakis et al.[2]	Mesh features, upper/lower word profiles	
Yue et al. [14, 42]	Stroke densities	Euclidean Distance Measure (EDM) and City Block Distance (CBD)
Al-Khatib et al. [34] and Shahab et al. [38]	Angular line, Concentric circle, 15-point DFT transform of vertical/horizontal profiles, Height, Width, Area and Aspect Ratios	
Rothfeder et al.[35]	Corner points	SLH, EDM, Corner Correspondences
Keaton et al.[18]	Projection profiles, cavity features	Minkowski distance (4 -norm), K-NN
Kotcz et al.[15]	Upper/lower word profiles, background-to-ink transitions	CBD, DTW
Adamek et al.[4]	Multi-scale convexities/concavities	DTW, K-NN (k =1)
Srihari et al.[8, 9, 46], Zhang et al. [41]	Gradient, Structure and Concavity features (GSC binary features)	Correlation Similarity Measure (CORR)*
Chen et al.[21, 28]	Upper/lower contours, background-to-ink transition (auto-correlation)	1D Hidden Markov Model
Kuo et al. [20, 24]	Pixel's value, Transactional information, Pixel's relevant position in row and column	Pseudo 2D Hidden Markov Model
Christophe et al. [53]	Image column height, Neighborhood size	NSHP-Hidden Markov Model
Ntzios et al. [5]	Open/closed concavity features	Binary Classification Tree
Tan et al. [27], Spitz et al. [26], Yue et al. [43], Linlin et al. [7]	Shape codes: derived by using other features such as Ascenders and Descenders	String Matching

Rothfeder et al. [35] have used an approach in which points of interest in the template and candidate images are initially determined, and then similarities between those points are recovered using correlation. The number of such similarities and the relative locations of the corresponding points are used to calculate the matching cost between the two images. To measure similarity between the two images, a voting method

* Redefined as dissimilarity measure

is used for counting the number of recovered correspondences. Every pair of features points whose similarity is greater than a given threshold is considered as a vote. The total number of votes received by each image is used to rank the retrieved images.

The Dynamic Time Warping algorithm is used to find the similarity between two images by aligning and comparing the sequences of feature vectors represented as time series. The time series are created by extracting a fixed number of features for every image column where each column represents the time axis in a horizontal direction. For each type of feature, a separate time series is created. These time series are then aligned and compared either jointly as a multivariate feature vector [1] or one by one separately [33]. The Dynamic Time Warping algorithm (DTW) has proven to be more successful for word spotting in handwritten and degraded documents [1, 29, 32].

Srihari et al. [8, 9, 46] and Zhang et al. [41] have used the well-known Correlation Similarity Measure (CORR) defined by Tubbs [31]. However, instead of using CORR directly, they have made some modifications and redefined it as a dissimilarity measure [78]. The more detailed and practical implementation of CORR has been presented in Chapter 2.

Group 2: Mapping Shape Codes: In this group, we have included those matching methods in which a word image is encoded into a relatively smaller set of predefined symbols, which is easier to recognize as compared to the original character set. For defining shape codes, the word image is initially divided into three zones: ascenders-zone, x-zone and descenders-zone, by estimating the base-line, x-line, mid-line, top-line and bottom-line, as shown in Figure 1.2. Generally, three different schemes for shape

coding have been reported, based on vertical bar patterns or character-level shapes or strokes.

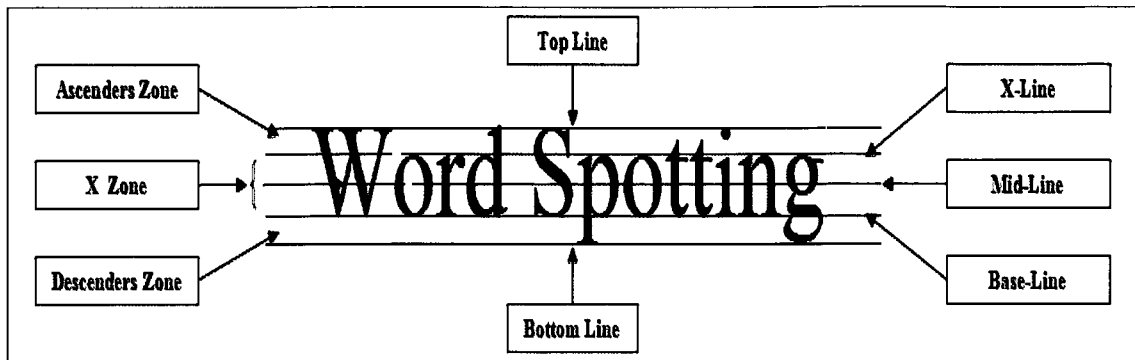


Figure 1.2: Illustration of important regions and lines used for Shape Codes estimation.

The pioneering work for keyword spotting in document images, using character shape codes, was presented by Tanaka et al. [22]. Their proposed system, called Transmedia Machine, identifies and encodes word-level components using two bits per character. The higher order bit encodes the presence of an ascender or descender, whereas the lower order bit encodes the frequency of crossing the mid-line by the corresponding character. The performance of this system improves when ascenders and descenders are treated separately by extending the original code. The Keyword is encoded in the same way and a search is initiated into coded documents for spotting other instances of the given query word.

Spitz et al. [26] have also used a similar but more robust approach, in which document images are transformed into character shape codes. These codes are then aggregated into word shape tokens and the tokenized texts are stored in a look-up table, i.e., an ASCII file. A query word, encoded in a similar way, is mapped to the look-up table using string matching techniques. The use of word shape tokens instead of actual words enables this method to be efficiently used for indexing and retrieval of scanned documents. A total number of six codes have been defined based on the features:

ascenders, descenders, number of connected components and deep-eastward concavity. These codes are used to classify the characters. This approach depends on the lexicon and segmentation of character cells for defining the character shape codes.

Tan et al. [27] have proposed a shape coding scheme based on vertical bars. In this approach, local extrema points, i.e., local minimum/maximum pixels, located on the word's contour are paired up and different vertical bar patterns are extracted. These vertical bars are classified into three groups, based on the presence of ascenders and descenders.

Lu et al. [43] have adopted a stroke-based approach for defining 29 word shape codes, which is based on straight strokes and traversal features. The straight strokes include vertical stroke lines, left-down diagonal lines and right-down diagonal lines. These strokes are extracted by using a run-length-based method. Whereas the traversal features represent the background-to-ink transitions, which are computed by scanning the word image column by column. These features are used to extract a code string called Left-to-Right Primitive String (LRPS) as it is sequenced from leftmost-to-right-most of a word, whereby a primitive is described by the ascender/descender features plus the shape of the stroke. Each character is given a standard Primitive String Token (PST). These tokens are stored in a look up table. A query word is converted into a code string in a similar way and the string matching technique is used to spot other instances of the given keyword. In this approach, partial phrases are also located by using a dynamic programming technique called inexact feature matching.

Another stroke-based shape coding method for word spotting has been presented by Linlin et al. [7], where the presence of straight/non-straight vertical strokes, ascenders/descenders and the total number of components are all used for encoding word images into code strings. Word images are decomposed into strokes, and each stroke is

assigned a shape code. For decomposition of word images into strokes the pixels lying on the mid-line are considered as decision points. If a pixel at this point, within an image column, is related to the background, then the entire column is converted to the background, otherwise the entire column is retained intact. There are eight codes used in total, which is a small amount compared to [43] where 29 complex codes are used. The instances of a query word are spotted using string matching techniques. This method of word spotting has been reported to be 20-40 times faster than OCR techniques [7].

Group 3: Training-based Matching: The third group, include those matching methods in which a so-called pre-trained model is used for keyword spotting. The well-known matching methods used for training-based word spotting include Hidden Markov Model (HMM) [12, 24, 53], K-Nearest Neighbor (K-NN) [4, 18] and Binary Decision Tree [5].

Chen et al. [21, 28] presented an approach for keyword spotting in printed documents which is based on Hidden Markov Model. Appropriate context-dependent character HMMs and sub-character HMMs are used for creating keyword HMMs, and non-keyword HMMs, respectively. Based on the two models, an *HMM network* is created, which is then searched by using Viterbi decoding on it, to spot the given keywords. Chen et al. [12] have modified the approach described in [28] by enabling it to spot phrases/partial words in document images. Unlike [28, 21] where up to four context-dependent models for each character are required, based on ascenders and descenders within a word, this new approach [12] uses a single HMM model for each character.

Kuo et al. [20, 24] have used pseudo 2D Hidden Markov Models, an extension to the standard HMM, which analyzes the word image in both horizontal as well as vertical directions. The model is called pseudo 2D because in order to avoid exponential

complexity it is not fully connected [48]. Two models are created, of which one models the keywords while the other models all the extraneous words, i.e., words other than the keyword. The search space is reduced by eliminating the unlikely words, i.e., any word that is not a keyword. Words are spotted by using the doubly embedded Viterbi decoding algorithm on the two HMMs.

Christophe [53] has used a hybrid model called Non-Symmetric Half Plane HMM (NSHP-HMM). This model combines HMM and Markov Random Field (MRF) where MRF is a second type of 2D-HMM with reduced connectivity [48]. At the HMM level, the binary patterns are analyzed column-by-column, whereas at the MRF level they are analyzed pixel-by-pixel.

1.3.4. Matching Levels

In the literature, image matching has been conducted at two levels for word spotting, i.e., whole word matching [1, 33, 46] and partial word matching [12, 20, 53], as shown in Table 1.6. In the first type, the global features of the whole word have to be matched with that of the template. In other words, the entire target word has to be similar to the template, based on some threshold value. Otherwise it will be rejected.

Table 1.6: Matching levels adopted for word spotting by various researchers.

Publications	Matching Level
Rath et al.[1], Jain et al.[33], Jawahar et al. [51], Balasubramanian et al.[52], Manmatha et al.[19], Rothfeder et al.[35], Cao et al. [3], Keaton et al.[18], Kotcz et al.[15], Adamek et al.[4], Srihari et al.[8, 9, 46], Zhang et al. [41], Al-Khatib et al.[34], Shahab et al. [38], Chen et al.[21]	Whole Word
Chen et al.[12, 28] , Kuo et al. [20, 24], Cho et al. [48], Christophe et al. [53], DeCurtins et al. [23]	Partial Word / Phrase

In the second type, not only do the exact instances of the given template have to be detected and located, but also the variants if any, have to be located, i.e., partial instances. For example, given a template word ‘*Science*’, the possible search results might be: *Science, Scientific, Scientist, Scientifically, etc.*

1.4. System Performance and Evaluation

The performance of a word spotting system is evaluated by using Precision-Recall Methods adopted from Information Retrieval literature. *Precision* is a measure of the degree of exactness, i.e., how much the results are exact. It can be calculated as follows:

$$\mathbf{Precision} = \frac{\mathbf{Number\ of\ relevant\ instances\ retrieved}}{\mathbf{Total\ number\ of\ instances\ retrieved}}$$

On the other hand, *Recall* represents the measure of completeness of a word spotting system. It can be calculated as follows:

$$\mathbf{Recall} = \frac{\mathbf{Number\ of\ relevant\ instances\ retrieved}}{\mathbf{Total\ number\ of\ instances\ present}}$$

The overall performance of a word spotting system can be described with the help of a graph called P-R-graph [6, 9]. This is drawn between the values of Precision and Recall measures. In addition, a software program called *trec-eval*, available on the internet [76], can also be used for evaluation of results [1]. In Table 1.7, we have presented the good precision rates achieved for each type of scripts in which word spotting techniques have been applied. It is important to note that the results shown in Table 1.7 have been achieved on different data sets, except for [1] and [35] which have used the same data set.

Table 1.7: Leading Precision Rates achieved in off-line word spotting Of various scripts.

Publications	Average Precision	Characteristics of Data Sets			
		Size	Variability & Quality of Data	Scripts	
Rath et al.[1]	65.34 %	2381	Handwriting of a single writer. Extracted from degraded and historical documents. Scanned from microfilms of George Washington's manuscripts.	Latin (Offline)	
Rothfeder et al.[35]	62.57 %				
Zhang et al. [41]*	92.18%	9312	Handwriting of 776 individuals. The test database contains only four different words, thus 2304 instances for each word. The quality of data is good.		
Kuo et al.[20]	99.00 %	26000	Printed in a single font size. A synthesized database with optimal quality.		
Kuo et al.[20]	96.00 %	26000	Printed in different font sizes. A synthesized database with optimal quality.		
Linlin et al.[7]	96.22 %	1845	Printed in different font sizes from 10 to 24 points and four types of font styles. Data is of good quality.		
Chen et al.[21]	96.00 %	2100	Printed in 8 different fonts including serif and sans-serif. Extracted from a text passage and tables of contents from five journals and conference proceedings. Data is of good quality.		
Jawahar et al.[51]	95.89 %	2507	Printed in various font sizes. Extracted from scanned pages of English books stored at the Digital Library of India (DLI). Images are of good quality.		
Jain et al. [33]	92.3 %	6672	Handwritings of 10 different writers. A digital device, i.e., Cross Pad has been used to capture the handwritten data.		Latin (Online)
Jawahar et al.[51]	94.51 %	2547	Printed in various font sizes. Extracted from scanned pages of an Amharic newspaper called Addis Zemen. Images are of good quality.		Amharic (Offline)
Jawahar et al.[51]	92.62 %	3354	Printed in various font sizes. Extracted from scanned Hindi documents archived at the Digital Library of India (DLI). Images are of good quality.	Devanagari (Offline)	
Park et al.[50]	89.84 %	1600	Printed in different font sizes ranging from 8 to 10 points. Extracted from 8 document images, printed in Korean. The quality of images is good.	Chinese (Offline)	
Yue et al. [42]	84.34 %	128	Printed in different font sizes and font styles. Extracted from various newspapers, printed in both traditional as well as simplified Chinese. Images are of good quality.		
Saabni et al. [85]	85 %**	8000	Extracted from Handwritten documents written by five different writers, Printed documents in different fonts and historical documents. The data is of reasonable quality.	Arabic (Offline)	

* The precision shown is for the top 100 matches only. The authors have presented the precision rates for 12 different top matches.

** Average precision for the three categories of data used in experiments.

Srihari et al.[8]	70%	20000	Handwritings of 10 writers. A synthesized database with good quality and accompanied with ground truth information.	Arabic (Offline)
Ntzios et al.[5]	89.06 %	12332	Handwritten characters and ligatures extracted from various historical manuscripts. Data is of poor quality, which is manually labeled and accompanied with ground truth information.	Greek (Offline)

1.5. Discussions: Issues and Future Work

As discussed earlier, Word Spotting techniques are more flexible when applied in the domains where documents are degraded and the vocabulary is unlimited and/or unseen. Nevertheless, there are some issues which affect the overall performance of a word spotting system. The precision rate is directly associated with computational complexity, i.e., achieving a higher precision rate involves a high computational cost. The precision rates shown in Table 1.7 are all based on limited sets of data and none of them have been applied to a huge amount of documents.

Other factors which affect the precision rate are variability of data and noises. In handwriting documents, the variability may be due to different writers, whereas in machine printed documents it can be caused by various font sizes and styles. From the work of Kuo et al.[20], it is evident that the precision rate for documents printed in a single font is higher than the rate for documents printed in many different fonts. Noises can be caused by degradation of documents, scanning sources, and most importantly, the page segmentation processes. The accuracy of the segmentation algorithms directly affects the precision rate of a word spotting system. The word spotting techniques which are not based on page segmentation can avoid the problems of improper segmentation. However, in segmentation-free methods accuracy may be affected by the methods used

for correctly determining zones of interest and by estimation of bounding boxes of the target images.

In addition, the sizes of the matching images may also affect the precision rate. For example, Zhang et al.[41], Leydier et al.[6] and Rothfeder et al.[35] have found that image lengths have a direct impact on the accuracy rate. They have reported that for longer images the retrieval precision is higher than that for shorter words. The main reason for this is that in shorter images, less information is available for reliable matching and a little variation can lead to misclassification. Image matching using Dynamic Time Warping has the ability to deal with these variations, however, it is associated with a high computational cost, i.e., it takes a long time to process large amounts of data.

To cope with the various problems discussed above, there is a need for more research to find efficient ways of image matching, page segmentation as well as to discover more effective features. Furthermore, a generalized model is required, which can be adopted for word spotting in any type of scripts. As discussed in section 1.2.1, several researchers have applied their proposed methods to different scripts, e.g., Leydier et al. [6, 10], Srihari et al. [46], and Jawahar et al. [51], etc. They have shown that the performance of the same word spotting system used for the various scripts is very similar. From their work, we can derive the hypothesis that: “It may be possible to design and develop a single and standard Word Spotting System which can be applied to almost all types of scripts and can achieve nearly equal performance for each type of scripts”.

1.6. Conclusion

Word spotting is an emerging technology and has various useful applications such as document indexing, information retrieval and information filtering. This technology is

still in the experimental stages and has been applied to limited sets of documents. It has produced good results for retrieving degraded, multilingual and complex documents as compared to OCR techniques. Word spotting in typewritten documents with good quality has given satisfactory results and can be commercially applied as an alternative to OCR systems. In the domain of degraded and historical documents, printed as well as handwritten, Word Spotting has achieved a higher accuracy rate as compared to OCR techniques. However, it has not achieved a precision rate acceptable for commercial applications. More efforts are needed to achieve a higher accuracy rate, improved performance and reduced computational cost. Robust and reliable page segmentation methods, proper feature selection and efficient image matching algorithms are required to take word spotting in handwriting documents to a commercial level.

CHAPTER 2

A New Word Spotting Approach for Word Image Retrieval from Gray-Scale Handwritten Databases

2.1. Introduction

As described in the preceding chapter, *Word Spotting* is a rich area of research, especially in the domains of degraded and historical documents. It is considered to be a complementary technique to Optical Character Recognition for document retrieval and analysis. In word spotting techniques, the documents are not converted into machine readable codes, as required for OCR-based techniques. Instead, the various patterns/images are compared based on their visual similarities. The important aspects/components of a word spotting system include *Page Segmentation Methods*, and *Image Matching Algorithms*.

Image matching is considered to be the most important as well as the more complex phase in word spotting technologies. The overall precision and performance of a word spotting system directly depends upon the Image matching method/algorithm. The effective image matching algorithm should be able to handle large amounts of data with greater speed and reliability. In the word spotting literature, various image matching methods have been reported. The most widely used ones are the Hidden Markov Model (HMM) [20, 21, 24, 28], Dynamic Time Warping (DTW) [1, 33, 51, 52] and Correlation Similarity Measure (CORR) [8, 9, 41, 46].

The HMM and DTW are the image matching methods which can be applied to both Gray-scale as well as Binary images. Both HMM and DTW can handle handwriting

variations very successfully, however, they have some disadvantages which disable them in certain domains. For example, HMM is a training-based matching method, that is, a training phase is required which is very time consuming. Due to this fact, in the domains where the data sets are large and/or are frequently changed, the use of HMM is not a good choice.

The Dynamic Time Warping (DTW) algorithm is a training-free matching algorithm and has been used for word spotting in degraded and historical handwritten documents [1]. This algorithm can effectively handle handwriting variations, but unfortunately it is very slow. Hence, DTW is not practical for handling large amounts of data, especially when they exceed a few thousand images.

The Correlation Similarity Measure (CORR) is a training-free algorithm, defined for matching two binary patterns [31]. It can handle a large amount of data with a greater speed and is also effective when there are large variations among the data. However, we can not compare the two feature vectors extracted from Gray-Scale images via the CORR algorithm. The images need to be binarized before matching. Due to this fact, this algorithm can not be used in the domains where the documents are degraded and/or historical.

In this chapter, we are presenting a new approach which enables the matching of the Gray-scale images using the Correlation Similarity Measure (CORR) algorithm. In this approach, the values within the feature vectors extracted from the Gray-scale images are converted into their binary equivalents and stored in the form of binary vectors. These binary vectors are then compared via the CORR algorithm.

The proposed approach has been tested on the newly created CENPARMI handwritten Pashto and Dari databases. These databases contain thousands of pre-extracted handwritten words of different classes, written in the corresponding languages. Pashto and Dari are well-known Indo-Iranian languages and are written in modified Arabic scripts. The two languages and their corresponding databases will be described in detail in Chapter 3.

The rest of this chapter is organized as follows. Section 2.2 provides background concepts and the related research work. Section 2.3 describes our approach in detail and in Section 2.4, the data set used in the experiments is described. We analyze the results of our experiments in Section 2.5. Some useful discussions and possible future works are presented in Section 2.6, whereas in Section 2.7 the concluding remarks are given.

2.2. Background

The Correlation Similarity Measure (CORR) is one of the eight Similarity Measures defined by Tubbs [31], for matching two binary patterns/images. CORR has been successfully applied by some researchers to match binary images in their word spotting systems [8, 9, 41, 46].

Srihari et al. [46] presented an interactive system called CEDARABIC based on the Correlation Similarity Measure (CORR) algorithm [31]. Their system is used for word spotting in three different types of scripts, i.e., Arabic, Latin and Devanagari. The working mechanism of their system is like a dictionary look-up. The user enters a query word, typed in English text, and the system finds its equivalent handwritten words in the pre-segmented document database. The users can also enter word images as queries, as an alternative to English texts.

Zhang et al. [41] have used the Correlation Similarity Measure (CORR) algorithm for image retrieval from a database of handwritten English words. They have applied the Dynamic Time Warping (DTW) on the same data set and have reported that the CORR algorithm is more effective and efficient than the DTW algorithm.

Srihari et al. [46] and Zhang et al. [41] have used the GSC binary features, which are the combination of three features, i.e., Gradient, Structural and Concavity features. These features measure the characteristics of an image at local, intermediate and global ranges, respectively. In their approach, an image is divided into 8 vertical zones and 4 horizontal zones, i.e., 4×8 sub-regions. Initially, the word image is vertically divided in such a way that each sub-region contains the same number of black pixels. Afterwards, it is horizontally divided in the similar way. Due to this division schemes, the resulting zones/sub-regions do not have equal heights and widths, i.e., have variable sizes. Hence, some of the zones may have more height but less width and vice versa. Once the word image is divided in the above mentioned manner, the GSC binary features are extracted from each zone/sub-region, i.e., 12 Gradient features, 12 Structural features and 8 Concavity features. The resulting binary vector of each word image consists of 1024 bits: 384 bits represent the Gradient features ($12 \times 4 \times 8$), 384 bits represent the Structural features ($12 \times 4 \times 8$) and 256 bits represent the Concavity features ($8 \times 4 \times 8$).

For extracting the GSC binary features, the word images need to be binarized, skew corrected and slant normalized [86]. But binarization of documents can lead to the loss of very useful information. In the domain of degraded and historical documents, binarization can wash out most of the important information. In addition, these features are not very tolerant to the variations found in different handwritings and can lead to a

high error rates when there are more variations among the data. It is not only difficult but in some cases impossible to detect the slant/skew angles in the cursive Arabic and modified Arabic scripts, especially Pashto scripts, at the word level and/or character level. Also, slant removal may distort the original shape of the target word image and lose some important information.

The success of an image matching method is closely associated with the selection of features and the manner in which these features are matched. The features which provide more detailed and reliable information about the various patterns are considered to be the good features. In addition, the useful features are those which are more tolerant to data variations, especially handwritings produced by many writers or various font sizes and styles. The need for such features is greater for degraded and historical documents, handwritten as well as printed.

Rath et al. [1] have described the various types of features for word spotting in historical and degraded documents, i.e., profile and transitional features. These features can provide effective information about the external shape as well as the internal structure of the word images. The profile and transitional features can play an important role in the domain of degraded and historical documents because they can be extracted from both Gray-Scale and Binary images. Furthermore, these features can also tolerate variations among the data; hence slant correction of the words and/or characters is not necessary.

Nevertheless, the problem is that we can not compare profile and transitional features, in their original forms, via the CORR algorithm. The CORR algorithm can only

match two binary vectors [31], that is, the feature values can either be 0 or 1. To handle this problem, we propose a new approach which is described in the following section.

2.3. Our Approach

In our approach, we have used the transitional and profile features, i.e., projection profile and word profile features. We have selected these features because, as mentioned earlier, they provide more detailed information about the internal structure and external shape of the word images. In the following subsections, we will describe our method of extracting these features from Gray-scale images and converting them into binary vectors, which enable their matching through the CORR algorithm.

2.3.1. Extraction of Features

The various features we used are divided into two sets, i.e., Feature Set 1 and Feature Set 2. For extracting Feature Set 1, every Gray-scale word image is initially divided into 100 zones, each one of size 10*10 pixels, as shown in Figure 2.1. From each zone, six types of features are extracted:

- i. Column-wise Projection Profile
- ii. Row-wise Projection Profile
- iii. Left Diagonal Projection Profile
- iv. Right Diagonal Projection Profile
- v. Column-wise Background-to-Ink Transitions
- vi. Row-wise Background-to-Ink Transitions

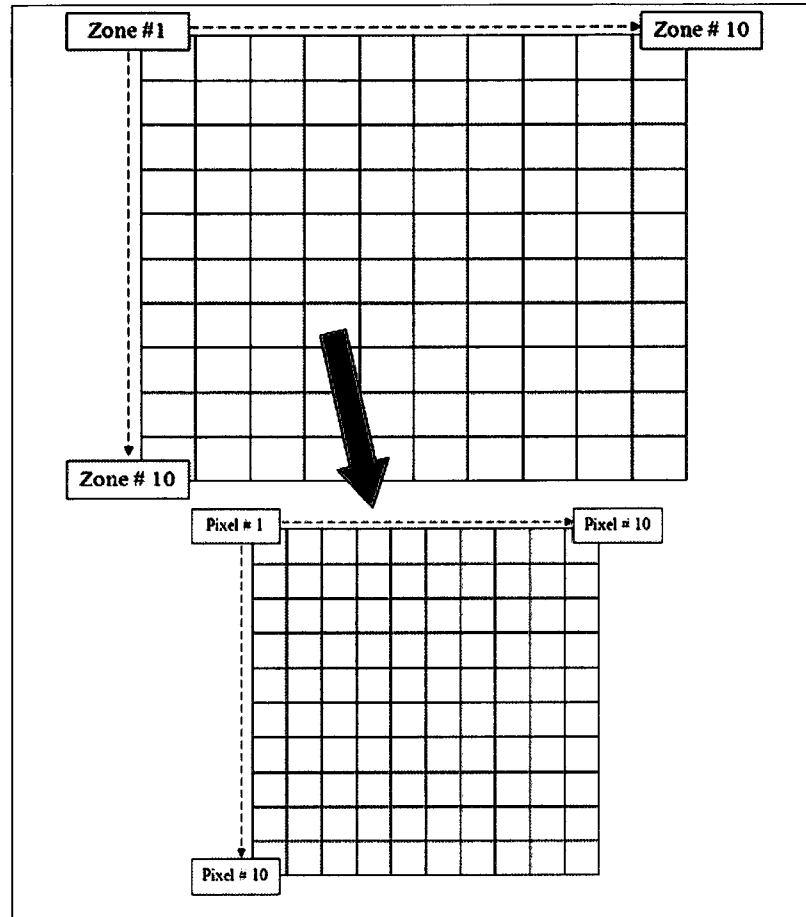


Figure 2.1: Dividing the word image into 100 zones, each one of size 10 * 10 pixels

Feature Set 2 consists of those features which are extracted from the entire image's columns/rows, i.e., the image is not divided into zones. The values of these features are divided by a threshold in order to make them consistent with the values of features in Feature Set 1. The four features included in Feature Set 2 are:

- i. Upper Word Profile
- ii. Lower Word Profile
- iii. Left Word Profile
- iv. Right Word Profile

The diagonal projection profile features, in Features Set 1, play a very significant role in the domain of documents written in Arabic and/or modified Arabic scripts. Most of the words written in Arabic or modified Arabic scripts contain either right/left diagonal or both left and right diagonal lines. Figure 2.2.a shows a Pashto word with both right and left diagonal lines, whereas Figure 2.2.b shows another Pashto word having three right diagonal lines.



Figure 2.2.a: A Pashto word having Left and Right diagonal lines.



Figure 2.2.b: A Pashto word having several Right diagonal lines.

It is worth mentioning that the Left/Right Diagonal Projection Profiles are extracted from the diagonal of each zone with a Radius = 5. The shaded regions in Figure 2.3.a and Figure 2.3.b show the Left and Right diagonal areas, respectively, of a zone from which we extract the projection profile features.

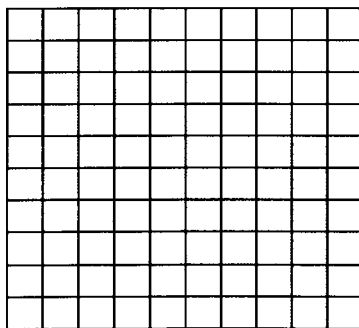


Figure 2.3.a: Left diagonal area of a zone with radius $R=5$.

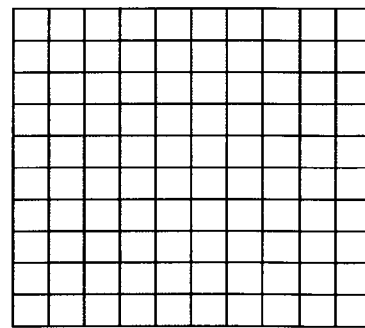


Figure 2.3.b: Right diagonal area of a zone with radius $R=5$.

All of the feature vectors of Set 1 and Set 2 are merged together into a 6400-valued feature vector. In the next subsection, we will describe how these feature vectors are converted into their binary equivalent vectors.

2.3.2. Conversion of Features to Binary Equivalents

The values of the feature vectors extracted from Gray-scale images are then converted into their binary equivalents. Each decimal value is represented by its four binary equivalents, as shown in Figure 2.4. Consequently, we get a large binary vector of 25,600 bits (i.e., $6400 * 4$) for every word image, where each set of four consecutive bits represent a single feature value. Since the values in the original feature vector ranges from 0 to 10, we use the right-most four bits of their binary equivalents and ignore the four extra zeros at the left.

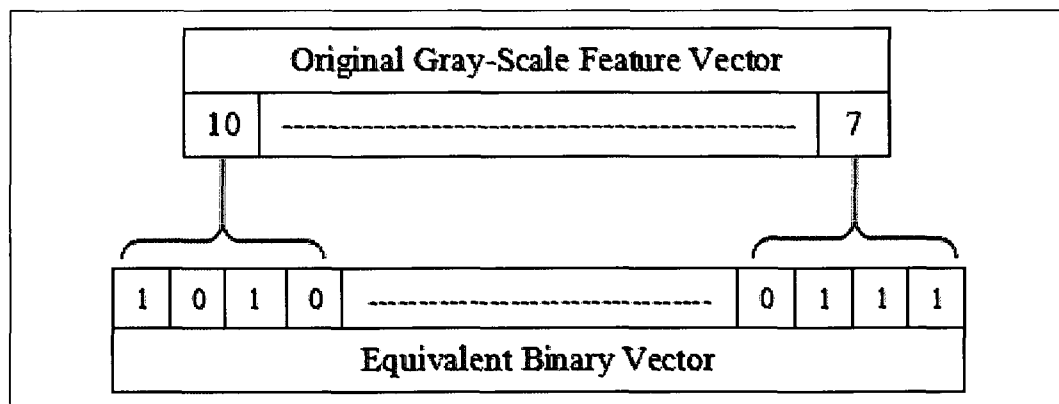


Figure 2.4: Representing decimal values of a Gray-Scale Feature Vector by their four binary equivalents.

Table 2.1 shows the original binary equivalents and the bits used in our experiments. The number of these bits can be increased or decreased based on the ranges of values in the Gray-scale feature vector. The binary equivalents of values lying between 0 and 255 are given in Appendix A.

Table 2.1: Binary equivalents of the decimal values used in our experiments.

Decimal Values	Binary Equivalents	
	Total Bits	Bits Used
0	0000 0000	0000
1	0000 0001	0001
2	0000 0010	0010
3	0000 0011	0011
4	0000 0100	0100
5	0000 0101	0101
6	0000 0110	0110
7	0000 1000	0111
8	0000 1000	1000
9	0000 1001	1001
10	0000 1010	1010

The binary vectors of all the target images are stored in the form of a vectors' database, which we will call as the Database of Features. This database is stored permanently in the computer's secondary memory and is called upon whenever a query is triggered. Figure 2.5 shows a snapshot of this Database of Features.

Index	Image Name	Feature Vector
0001	PSH0069_P02.tif	1 0 1 1 ----- 0 1 0 1
⋮	⋮	⋮
4200	PSH0021_P02.tif	1 0 1 1 ----- 0 1 0 1

Figure 2.5: A snapshot of the Database of Features.

The advantage of this feature database is that each time a query is triggered; we do not have to extract features from the target images. If new images are added to the target data set, their features are extracted and merged to this database.

2.3.3. Image Matching

For image matching, we use the Correlation Similarity Measure (CORR) algorithm [31]. The similarity between the two binary patterns, i.e., binary vectors of the template and the target images, is defined as:

$$\text{Similarity}(\text{BV}_{\text{template}}, \text{BV}_{\text{target}}) = \frac{(\text{D}_{11} * \text{D}_{00}) (\text{D}_{10} * \text{D}_{01})}{\sqrt{S}} \quad (1)$$

where:

$$S = (\text{D}_{10} + \text{D}_{11}) (\text{D}_{01} + \text{D}_{00}) (\text{D}_{11} + \text{D}_{01}) (\text{D}_{00} + \text{D}_{10})$$

$$\text{D}_{00}(\text{BV}_{\text{template}}, \text{BV}_{\text{target}}) = \text{CBV}_{\text{template}} * \text{CBV}_{\text{target}}$$

$$\text{D}_{01}(\text{BV}_{\text{template}}, \text{BV}_{\text{target}}) = \text{CBV}_{\text{template}} * \text{BV}_{\text{target}}$$

$$\text{D}_{10}(\text{BV}_{\text{template}}, \text{BV}_{\text{target}}) = \text{BV}_{\text{template}} * \text{CBV}_{\text{target}}$$

$$\text{D}_{11}(\text{BV}_{\text{template}}, \text{BV}_{\text{target}}) = \text{BV}_{\text{template}} * \text{BV}_{\text{target}}$$

and:

$$\text{BV}_{\text{template}} = \text{Template Binary Vector}$$

$$\text{CBV}_{\text{template}} = \text{Complement of } \text{BV}_{\text{template}}$$

$$\text{BV}_{\text{target}} = \text{Target Binary Vector}$$

$$\text{CBV}_{\text{target}} = \text{Complement of } \text{BV}_{\text{target}}.$$

A target word image will be considered as an instance of the given template word if the following rule is satisfied:

$$[\text{BV}_{\text{template}} \equiv \text{BV}_{\text{target}}] \rightarrow [\text{Similarity}(\text{BV}_{\text{template}}, \text{BV}_{\text{target}}) \geq \text{Threshold}] \quad (\text{Rule 1})$$

According to Formula 1, the range of similarity between the two binary patterns is between -1 and 1. Yet, we have adopted the approach described by Zhang et al. [78]. According to this approach Formula 1 can be redefined as a dissimilarity measure in the

form of Formula 2. The range of the dissimilarity values calculated by Formula 2 is between 0 and 1.

$$\text{Dissimilarity}(BV_{\text{template}}, BV_{\text{target}}) = 1 - \frac{(D_{11} * D_{00}) (D_{10} * D_{01})}{\sqrt{S}} \quad (2)$$

In this case, a target word image will be considered to be an instance of the given template word if the following rule is satisfied:

$$[BV_{\text{template}} \equiv BV_{\text{target}}] \rightarrow [\text{Dissimilarity}(BV_{\text{template}}, BV_{\text{target}}) \leq \text{Threshold}] \quad (\text{Rule 2})$$

The probability of a target image to be an instance of the template image will be higher if the value of Formula 2 for these two images is smaller and vice versa. Consequently, we get a list of dissimilarities between the template and all the target images, as shown in Figure 2.6. The value of the ‘*Threshold*’ is actually the dissimilarity value at a position P in the dissimilarity list shown in Figure 2.6.

Position (P)	Dissimilarity Value
0001	0.1
0002	0.3
⋮	⋮
4200	0.9

Figure 2.6: A sample of the sorted dissimilarities list.

This list is sorted in ascending order, thus entries coming on the top of the list are those images which are more similar to the template and/or are instances of the given template. On the other hand, entries on the bottom are relating to those images which have the least similarity with the template image and can not be its instances. This list is used to retrieve the instances of the given template from the database of target word

images. It is created for each query and is discarded from the memory when the query is completed.

2.4. Data Sets

The proposed approach has been tested on the newly created CENPARMI handwritten Pashto and Dari databases, which will be described in Chapter 3. These databases contain thousands of handwritten words of different classes. The data used for the experiments has been taken from the Gray-scale versions of these databases. In the next section, we are going to describe the various aspects of the target data sets.

2.4.1. Variations in the Data

Variations among the target data play an important role as they have a direct impact on the performance of any Word Spotting system. We can divide the variations among the data into two types, i.e., *Inter-class word variations* and *Intra-class word variations*. The Inter-class word variations are those variations which are found among the words belonging to different word classes. Whereas, the Intra-class word variations are those variations which are found among the words belonging to the same word class. As far as the Inter-class word variations are concerned they are directly related to the precision rate, i.e., the more the Inter-class word variations the better the precision rate. This is because the word spotting techniques compare the two word images, i.e., a template word image and a target word image, based on their visual similarities. The words related to one class which show more variations with the words related to the other class, will have fewer visual similarities. Hence, there will be a very small chance of misclassification. In

Figure 2.7, we show Pashto words of 15 different classes which have very high Inter-class word variations.

ختم	تیکس	باقی	کرور	نهټ
کوکرام	ملی میټر	روه	افغانی	اتیا
اوه	زر	واجب الادا	خلوینت	زخیره

Figure 2.7: Inter-class word variations of 15 different Pashto words.

The increasing Inter-class word variations are not a problem, but rather an advantage. However, the decreasing Inter-class word variations are a critical problem, which always leads to a high error rate, i.e., a misclassification problem. Both Dari and Pashto scripts have many challenges due to fewer Inter-class word variations. In these scripts, many words have the same basic body and only the diacritic marks make them different from each other. A human reader can comprehend these words, however, in word spotting techniques, the diacritic marks may be lost during the pre-processing stage, especially for noise removal, which can create problems and can lead to misclassification.

In Figure 2.8, we show eight groups of word classes taken from the Pashto database. Each group consists of words related to different classes, which show very few Inter-class words variations, i.e., in each group the words within one class are very similar to the words within the other class. As we can see that the words of the classes within some groups are so similar that without careful reading, a native Pashto speaker can easily be deceived. For example, the two classes included in group 1 are so similar to each other that even a human reader can make mistakes in distinguishing the various words relating to the two classes. The problem becomes worse when the words of such

classes are distorted and/or of poor quality. Handling such kinds of problems requires more intensive research work and development efforts. Therefore, we have tried to avoid such problems by not using those classes of words which can be confused easily with each other. In the future, we will make efforts to cope with the problem of decreasing Inter-class word variations.

Group #	Inter-Class Word Similarities			
1	تقدار		مقدار	
	Number		Amount	
2	كلوگرام		ملوگرام	
	Kilogram		Milligram	
3	میترا	لیتر	قیز	
	Meter	Liter	Item	
4	شپتہ	نیتہ	بدیہ	
	Sixty	Period	Price	
5	قسم	مجر	ختم	
	Article	Volume	Expire	
6	سنتی میٹر	ملی میٹر	ملی لیٹر	
	Centimeter	Millimeter	Milliliter	
7	شش	س	س	
	Twenty	Hundred	Ten	
8	انچہ	پنچہ	نہوہ	دبہ
	Inch	Five	Nine	Cottons

Figure 2.8: Various classes of words which show more similarities with each other.

On the other hand, the Intra-class word variations, that is, variations that exist among the words belonging to the same class, are inversely related to the precision rate of

a word spotting system. In other words, the fewer the variations that can be found among words of the same class, the higher the precision rate and vice versa. The Intra-class word variations are considered to be a big problem in word spotting research. Many researchers have tried to avoid Intra-class word variations either by using the handwritings of a single writer [1, 35] or by using a very limited number of writers [8, 9, 33]. The variations can exist even in the handwriting of a single writer. In Figure 2.9, the Intra-class words' variations have been shown for a Pashto word written by 15 different writers.

کھاټه	کھاټه	کھاټه	کھاټه	کھاټه
کھاټه	کھاټه	کھاټه	کھاټه	کھاټه
کھاټه	کھاټه	کھاټه	کھاټه	کھاټه

Figure 2.9: Intra-class word variations among the various instances of a word.

These Intra-class word variations are not only a big problem in handwritten documents, but also in the domain of printed documents. In the latter case, the variations are caused by different font sizes and styles, etc. Kuo et al. [20] have shown that the precision rate for the words printed in single fonts is higher, i.e., 99%, than that for words printed in different font sizes, i.e., 96% precision.

2.4.2. Size of Test Data Sets

We have used two data sets of which the first one consists of 4,599 Pashto words of 21 different classes. The second data set consists of 4,599 Dari words of 21 different classes. The total number of words per data set is equal to the number of word classes multiplied by the number of writers. As shown in Table 2.2, there are 219 writers for both

Pashto and Dari databases and each person has written a word only once, hence each class of words has 219 different samples. All the word images have been stored together in a single directory, which will be called the Database of Target Word Images, in the remainder of this chapter. It is important to mention that words of the two languages have been stored and processed separately. In Section 2.5, we will present the results of various experiments based on a different number of writers, i.e., using different levels of Intra-class word variations.

Table 2.2: The handwritten data used in the word image retrieval experiments.

Language	Size of Data Sets used	Number of Word Classes	Number of Writers
Pashto	4,599 words	21	219
Dari	4,599 words	21	219

For both Pashto and Dari languages, we have used 4,578 word images, out of the 4,599 word images, as the Test Data Set, whereas 21 words have been used as the template words. Each word of the 21 template words represents a different word class. The template words were randomly selected from the entire Data Set of the corresponding languages. We have selected those words as the template words which were nicely written by the writer. As mentioned earlier, each word in our Data Sets has 219 instances written by 219 different writers. Hence, after selecting a word image as a template, the other 218 instances were included in the test data set.

As described in the preceding section, most of the words in our two databases show very little Inter-word variations. Handling the decreasing Inter-class word variations and increasing Intra-class word variations at the same time is a great challenge in word spotting. In this work, we have tried to avoid the problem of decreasing Inter-class word

variations by selecting only those Pashto and Dari words which show higher Inter-class word variations. For both the languages, 21 classes of words have maximum Inter-class word variations. We have focused more on the problem of Intra-class word variations by using the handwritings of 219 different writers. In the future, we will try to handle the Inter-class word variations as well as the Intra-class word variations simultaneously.

2.4.3. Pre-Processing of Data

The data of Pashto and Dari databases is of good quality and all the background noises have been removed by using a median filter of kernel size $3 * 3$, as will be described in Chapter 3. The pre-processing module in our experiments performs two important tasks, i.e., removing all the white margins from the word images and scaling all the word images to the same dimensions, as shown in Figure 2.10.

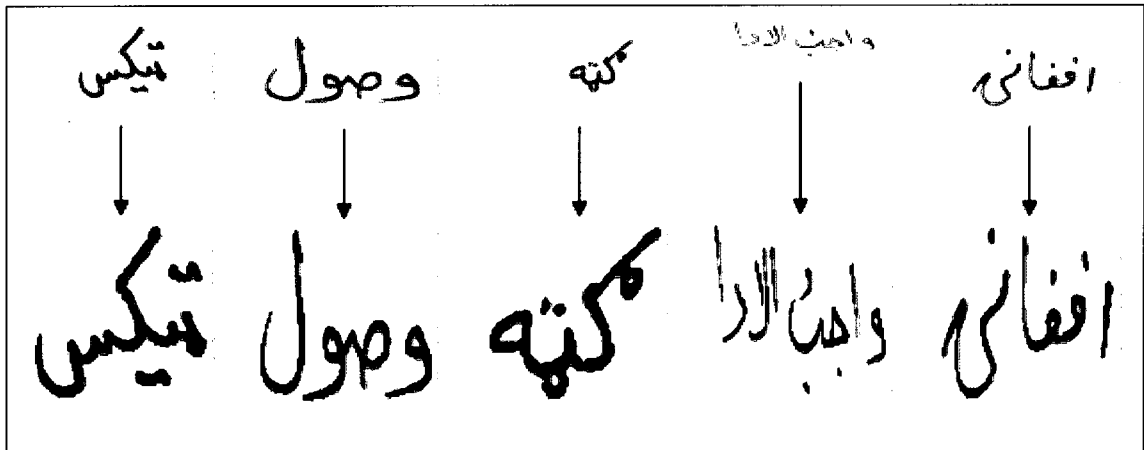


Figure 2.10: Pre-processed images with no white margins and all scaled to the same size.

We have tried various sizes for scaling, i.e., $50 * 50$, $64 * 64$ and $100 * 100$ pixels. After analyzing our results, we found that the images of size $100 * 100$ pixels produced the best results. Other researchers have also reported that the bigger the sizes of the word images, the higher the precision rate and vice versa [6, 35, 41]. The advantage of using

large image sizes is that more information can be extracted, which makes the matching process less prone to data variations, and ultimately leads to a higher precision rate.

2.5. Experiments

In this section, we present the application of our method for “Word Image Retrieval” from the two handwritten databases, i.e., Pashto and Dari words. We will give an overview of the various modules included in our experimental set up as well as will present the results of the various experiments conducted on the two data sets

2.5.1. Overview of Applications

In our experiments, a word image is selected as a template image and a search is initiated into the Database of Target Word Images to retrieve the other instances of the selected template word. We have evaluated the proposed approach by a series of nine experiments for both Pashto and Dari languages. In each experiment, a word was chosen as a template and its corresponding instances were called upon from the Database of Target Word Images. We can summarize the various steps in our experiments as follows:

- i. Extract features from the given template.*
- ii. Call upon the Database of Features: shown in Figure 2.5.*
- iii. Compare the binary vectors of all the target images with that of the template image, using the CORR algorithm.*
- iv. Create a dissimilarity list by sorting the results of **Step iii**, in ascending order: as shown in Figure 2.6.*
- v. Select the threshold value ‘TH’: The dissimilarity value at a Position P, as shown in Figure 2.6.*

- vi. Retrieve all those images from the target data set whose dissimilarity value is either equal to or smaller than the threshold 'TH'.

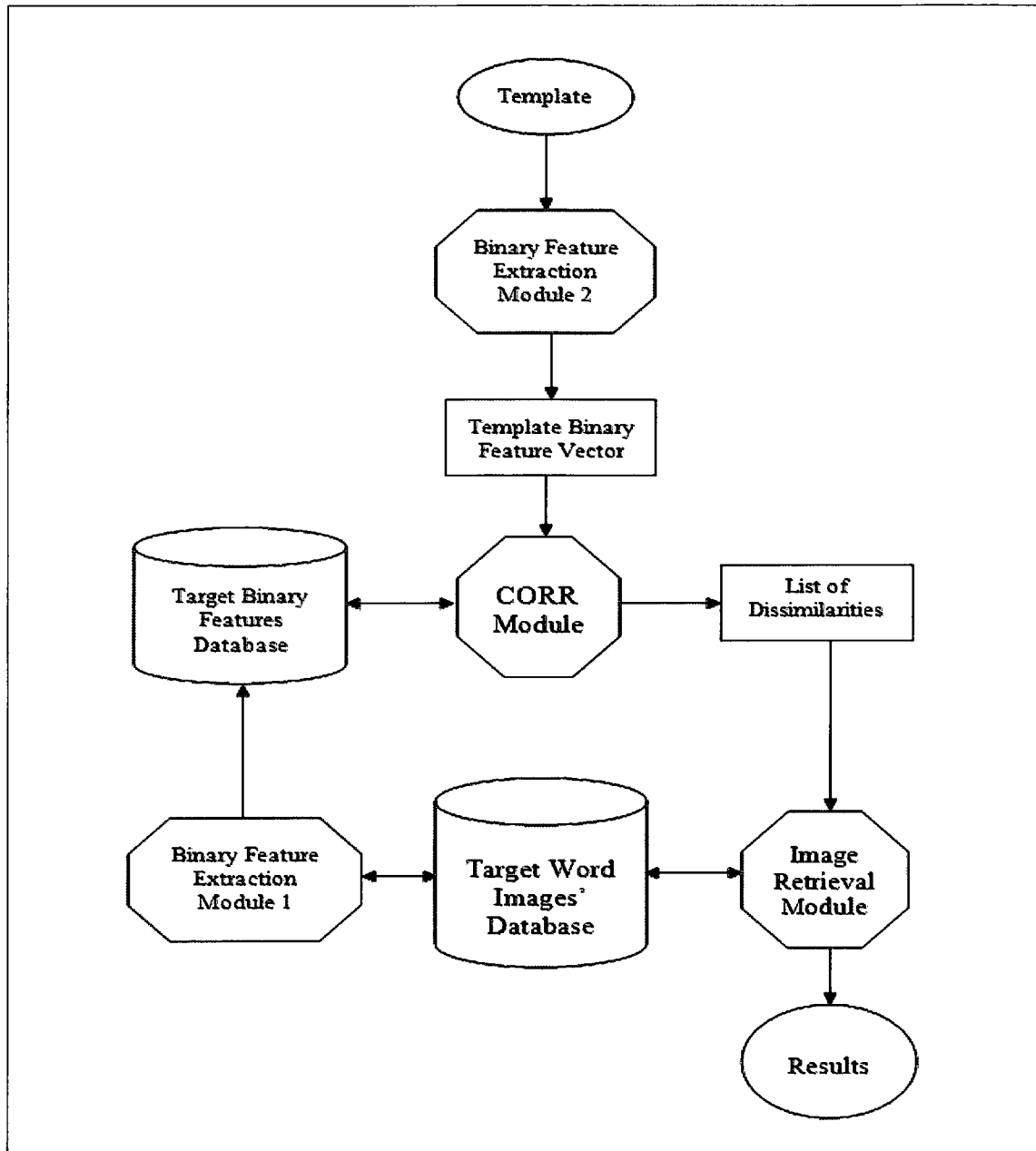


Figure 2.11: Applications' Overview

Figure 2.11 depicts the various modules used in our experiments. There is no difference between Binary Feature Extraction Module 1 (BFEM-1) and Binary Feature Extraction Module 2 (BFEM-2), with respect to programming logic. However, the

BFEM-1 is called upon whenever the Database of Target Word Images is updated, i.e., either some new images are added or the existing images are modified. On the other hand, the BFEM-2 is called upon each time a query is triggered. In addition, the feature vector created by the BFEM-2 is for the template/query image only and is temporarily stored in the primary memory. This feature vector is discarded once the query is executed, i.e., the instances of the given template are retrieved.

2.5.2. Analysis of Results

We have conducted eight experiments on both Pashto and Dari data sets. Table 2.3 shows the average precision rates for the nine experiments on Dari and Pashto words. For both Dari and Pashto data sets, each experiment has been repeated for 21 different template/query images, selected from the corresponding data sets.

Table 2.3: Average precision rates for nine experiments on Dari and Pashto words.

Experiment	TH at P =	Average Precision Rates	
		Dari	Pashto
1	50	99.87 %	99.52 %
2	80	99.83 %	99.23 %
3	100	99.13 %	98.24 %
4	120	98.28 %	97.38 %
5	140	96.24 %	95.71 %
6	160	94.08 %	92.95 %
7	180	90.74 %	89.63 %
8	200	86.6 %	85.31 %
9	218	80.96 %	80.63%
Average Total		93.75 %	93.18 %

In Table 2.3 the precision rate drops when the value of parameter P increases from experiment number 1 to experiment number 9. This shows that the precision rate drops when there are more handwriting variations, i.e., Intra-class word variations, increase.

Tables 2.4 and 2.5 show the precision rates for each individual query/template word of Dari and Pashto languages, respectively. In Tables 2.4 and 2.5, the precision rates decrease for most of the templates from left to right. However, we can observe different patterns of fluctuations in the precision rates for some of the templates. For example, initially the precision rate increases from left to right and then drops in the same direction, i.e., template # 16 in Table 2.5. Moreover, for some templates the precision rates fluctuate in the order *decrease-increase-decrease*, i.e., template # 3 in Table 2.4 and templates # 1, 11 and 19 in Table 2.5.

The fluctuations in the precision rates are due to the decreased Inter-class word variations problem, i.e., words belonging to other word classes have less dissimilarity with the template word as compared to the words of the template's own classes. As mentioned earlier, the word images are retrieved based on their dissimilarities which are sorted in ascending order. Hence, such erroneous words come first in the dissimilarity list and get priority over the real instances of the template during the image retrieval process. When the threshold value is increased, the real instances of the template get the chance of retrieval. Hence the precision rate increases instead of decreasing. As mentioned earlier, to avoid the decreased Inter-class word variations problem we have chosen only those word classes which show maximum inter-class word variations.

Table 2.4: Individual precision rates for the nine experiments conducted on Dari words.

Template Number	Individual Precision Rates for the Nine Experiments									Average Precision
	1	2	3	4	5	6	7	8	9	
1	100	100	98	96.67	95.71	93.75	90.56	84.5	79.73	93.21 %
2	100	100	100	99.17	95.71	92.5	91.67	84.5	79.73	93.7 %
3	100	100	99	99.17	98.57	96.25	92.78	88	78.83	94.73 %
4	100	100	99	95	92.14	88.75	87.78	84.5	79.73	91.88 %
5	100	100	100	100	97.14	93.75	88.33	84	77.03	93.36 %
6	100	100	100	100	98.57	98.13	93.89	92	86.49	96.56 %

7	100	100	96	100	100	100	99.44	98	93.24	98.52 %
8	100	100	100	93.33	90.71	85.63	81.67	77.5	72.52	89.04 %
9	100	100	100	100	99.29	99.38	99.44	99	94.59	99.08 %
10	100	100	99	98.33	97.86	98.13	95.56	92	87.84	96.52 %
11	98	98.75	98	99.17	91.43	85.63	81.67	77	72.07	89.08 %
12	100	98.75	100	98.33	97.86	96.88	93.89	90	84.23	95.55 %
13	100	100	100	100	100	100	96.11	91	86.04	97.02 %
14	100	100	99	97.5	95	93.13	82.78	76.5	72.07	90.66 %
15	100	100	99	97.5	93.57	89.38	85.57	80.5	73.87	91.04 %
16	100	100	98	96.67	95	91.88	89.44	83.5	78.99	92.61 %
17	100	100	100	97.5	95.71	91.88	91.67	84.5	78.99	93.36 %
18	100	100	99	94.17	92.14	85.63	81.11	79.5	74.43	89.55 %
19	100	100	100	97.5	95.71	94.38	91.11	88.5	82.65	94.43 %
20	100	98.75	99	95.83	93.57	90	87.22	81.5	78.99	91.65 %
21	100	100	100	100	99.29	98.13	96.67	93	88.13	97.25 %
Total Average Precision										93.75 %

Table 2.5: Individual Precision rates for the nine experiments conducted on Pashto words.

Template Number	Individual Precision Rates for the Nine Experiments									Average Precision
	1	2	3	4	5	6	7	8	9	
1	100	98.75	97	98.33	95	90	85.56	82.5	76.71	91.54 %
2	98	97.5	95	90.83	86.43	81.88	75	70	66.21	84.54 %
3	100	100	100	100	100	96.25	92.78	87.5	80.82	95.26 %
4	100	100	99	97.5	97.14	95.63	91.67	86	81.74	94.3 %
5	100	98.75	93	90.83	88.57	83.75	81.67	75.5	69.86	86.88 %
6	100	100	100	100	100	100	98.89	96.5	91.32	98.52 %
7	100	100	100	99.17	98.57	96.25	93.89	92	86.76	96.29 %
8	100	100	100	100	100	99.38	95.56	93	88.13	97.34 %
9	96	95	93	91.67	87.86	85	81.11	77	72.60	86.58 %
10	100	98.75	100	100	100	100	100	97.5	92.24	98.73 %
11	100	100	99	99.17	97.14	91.88	89.44	83.5	79.91	93.34 %
12	100	100	100	100	100	98.75	95.56	91.5	88.59	97.16 %
13	100	100	100	100	100	99.38	97.78	93	87.22	97.49 %
14	100	100	100	100	99.29	98.75	97.22	93.5	88.13	97.43 %
15	100	100	100	97.5	95	92.5	88.89	83	79.91	92.98 %
16	98	98.75	99	98.33	94.29	91.25	86.67	82.5	77.17	91.77 %
17	98	97.5	93	91.67	87.86	80.63	76.67	73	67.58	85.1 %
18	100	100	99	96.67	95.71	92.5	89.44	86.5	82.19	93.57 %
19	100	100	99	99.17	96.43	91.25	84.45	79.5	73.97	91.53 %
20	100	98.75	98	96.67	96.43	94.38	91.67	85.5	82.19	93.73 %
21	100	100	99	97.5	94.29	92.5	88.33	82.5	79.91	92.67 %
Total Average Precision										93.18 %

We have chosen not to compare our results to those of the other word spotting researchers as most of them have used different matching algorithms, applied on different datasets written in different types of scripts. However, Srihari et al. [8, 9, 46] have applied the CORR algorithm for word spotting in Arabic documents. Hence, to some extent we can compare our results to their results, as shown in Table 2.6. The precision rate, presented by Srihari et al. [8, 9], has been averaged for 150 queries, whereas in our case the precision rates have been averaged for 189 queries for Pashto and 189 queries for Dari language.

Table 2.6: Comparison of results based on various aspects.

Approach	Test Data Set		Queries	Average Precision	Features
	Size	Writers			
Srihari et al. [8, 9]	4,000 Arabic Words	2	150	70 %	GSC Binary Features
Ours	4,578 Pashto Words	218	189	93.18 %	Profile, Transitional
	4,578 Dari Words	218	189	93.75 %	

Srihari et al. [8, 9] have used 100 handwritten Arabic documents written by 10 writers, i.e., each writer has written 10 documents. According to the authors, these 100 documents contained 20,000 words, i.e., the set of every 10 documents written by each writer contained 2,000 words. The results shown in Table 2.6 were achieved when they used the documents written by only two writers as the Target/Test Data Set, i.e., 4,000 words. The rest of their documents, written by other eight writers, were used to provide the template words.

GSC binary features, used by Srihari et al. [8, 9, 46], were originally defined and used for hand-printed digits and character recognition [86]. These features can be properly extracted from isolated digits or characters. Hence, before the extraction of these features, boundaries of the individual characters/strokes, within a word, should be

defined/marked out [86]. In addition, for extraction of the Gradient features, only the direction of the gradient vector is considered and used as a feature value. Due to this reason, the images must be slant normalized and skew-corrected. For languages written in Latin scripts, these features may be useful, however, for Arabic scripts, which are written in a complex cursive manner, these features may not prove to be very effective.

The use of GSC binary features for word spotting, by Srihari et al. [8, 9, 46], was motivated by the successful use of these features for writer identification and verification in the area of forensic research [87, 88, 89]. Handwriting identification is considered to be an image retrieval task with an extra effort to determine authorship [41]. However, the image retrieval philosophy for analyzing the handwritten individuality is totally different from that of the word spotting. In the former, the search is narrowed down to retrieve the words/characters/digits written by a specific writer. In the word spotting, the search is initiated to retrieve all the probable instances of a given word/template, which may have been written by more than one writer or printed in more than one font size/style.

The use of GSC features for word spotting in documents written by one writer may prove to be successful, however, in the domains where there are more writers this approach can not work. Srihari et al. [8, 9, 46] have shown that when they used the handwritten words of five writers, as the Target/Test Data Set, the precision rate was 55 %. However, when they used the handwritten words of only two writers, as the Target/Test Data Set, the precision rate increased from 55 % to 70 %. Hence, it shows that GSC binary features are more prone to variations and are not successful in the domains where the number of writers of the target/test documents is greater than one.

On the other hand, each of our two data sets contains 4,599 words, written by 219 different writers. For each word class, we have used one representative word as a template word, i.e., handwriting of one* writer only, and all the words written by 218 different writers have been used as the target/test data set, i.e., 4,578 words. Hence, each of our data sets contains 109 times more variations, i.e., Intra-class word variations, as compared to Srihari et al. [8, 9]. As mentioned earlier, the Intra-class word variations, whether caused by different writers or different font sizes and/or font styles, is a big problem in word spotting. Our approach can handle this problem very successfully. In addition, our results are based on the Pashto and Dari languages, which are written in modified Arabic scripts. These scripts are more challenging, especially Pashto scripts, than the standard Arabic scripts.

Moreover, the approach used by Srihari et al. [8, 9, 46] can not be applied to Gray-scale images, especially degraded and historical, because for the extraction of GSC binary features the images need to be binarized. On the contrary, our approach can be applied equally to Gray-scale as well as binary images. Overall, we can argue that our approach is more flexible and robust as compared to Srihari et al. [8, 9].

As shown in Table 2.6 our system has given almost similar results for Pashto and Dari Data Sets, i.e., 93.18 % and 93.75 %, respectively. Both data sets contained different words, written by different writers. Furthermore, Pashto scripts are different from Dari scripts. Hence, it shows that our approach is consistent, i.e., performance is almost the same for both languages and is not affected by scripts' type. This supports our hypothesis presented in Chapter 1:

* It is important to mention that all the representative words, i.e., template words, for all the word classes have not been written by a single writer. The writer of the template word representing one class of word was different from that of the other class of word.

“It may be possible to design and develop a single and standard Word Spotting System which can be applied to almost all types of scripts and can achieve nearly equal performance for each type of scripts”.

We have used MATLAB 7.4.0 (R2007a) for programming purposes. All of our experiments were conducted on a computer system (i.e., HP Pavilion dv6000 with Microsoft windows Vista, home edition) having two processors of 1.6 Giga Hertz, and two Giga Bytes of Random Access Memory (i.e., 2 GB RAM). The average time taken for matching every set of two word images was about 0.00143 seconds.

2.6. Discussions and Future Work

As various kinds of features have been extracted and matching methods have been applied by researchers, efforts are being made to enable word spotting technology for commercial applications. To achieve this goal, a word spotting system should be more reliable and able to process a huge amount of data with greater speed and high precision.

Speed and accuracy are considered to be the two major concerns of the word spotting research community. There is always a trade-off between these two important factors, i.e., one is achieved at the cost of the other. For example, getting high accuracy requires the processing of more information about the various patterns, which results in a high computational cost, i.e., processing time taken by the image matching algorithm and memory cost.

Nowadays, due to the development of information technology, it is easy to handle the problem of too much memory. However, the processing speed of the matching algorithms is still a challenge for the word spotting community. The CORR is considered to be the faster image matching algorithm in the word spotting community [8, 9].

However, it was originally defined for matching binary patterns and could not be applied to Gray-scale images. We have emphasized the use of Gray-scale images because from these images we can extract more information. Binarization of images, especially the degraded and faded documents, leads to the loss of some useful information. The more detailed information available about the patterns the higher the accuracy of the system. Less information can be easily affected by variations among the data, i.e., handwriting variations or different font sizes/styles, which will ultimately lower the precision rate. Our approach has brought an end to the need of image binarization required for the CORR algorithm. Now, we can extract more detailed information from the Gray-scale images and match them with greater speed and high precision rate.

In the future, the proposed approach could be tested by replacing the Correlation Similarity Measure (CORR) with some of the other seven Similarity Measures defined by Tubbs [31], i.e., Jaccard and Needahm (JACCARD), Dice (DICE), and Yule (YULE). Some optimization techniques may be introduced to minimize the sizes of the large binary vectors. Moreover, other new features and pre-processing techniques could be applied.

The word spotting technique has been applied for the first time in Dari and Pashto scripts. The results show that the proposed approach is effective for word spotting in these challenging scripts. In this work, we have used pre-extracted words, however, in the future, we could apply this method to search and locate the key words in full text handwritten documents. Moreover, the same method can be applied to other languages written in Arabic scripts, as well as other types of scripts, such as Latin, Greek, Chinese and Devanagari, etc.

2.7. Conclusion

A new approach has been introduced, in which the Correlation Similarity Measure Algorithm (CORR) has been used for matching Gray-scale images. The detailed information in the form of profile and transitional features extracted from Gray-scale word images have been successfully matched with greater speed and higher accuracy. Our approach has now made it possible to apply the CORR algorithm to Binary as well as Gray-Scale images, especially to historical and degraded documents.

CHAPTER 3

The Handwritten Databases of Pashto and Dari Languages

3.1. Introduction

In this chapter, we will present two large new handwritten databases for Pashto and Dari languages. These databases are the results of the efforts made at the Center of Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, to create standard, multipurpose handwritten databases of Pashto and Dari languages. The creation of these databases is an important step towards the Optical Character Recognition (OCR) of Pashto and Dari languages, as the availability of a standard and useful database always plays a central role in the recognition of any language. The robustness of any recognition or segmentation algorithm directly depends upon the data set used for experimentation and evaluation.

Nowadays, many people and organizations are interested in the Optical Character Recognition of the Indo-Iranian languages written in Arabic scripts. There are five well-known Indo-Iranian languages written in Arabic and/or modified Arabic scripts: Arabic, Farsi, Urdu, Dari and Pashto languages. Some useful databases have been created for some of these languages over the last several years. For example, Solimanpour et al.[79] presented a handwritten collection of data about Farsi isolated digits, English isolated digits, isolated alphabets, Farsi dates and Farsi legal amounts. This database is useful for handwritten check recognition research. The ‘CENPARMI Arabic cheques’, presented by Al-Ohali et al.[80], is a large database. This database contains Arabic hand printed bank cheques and is therefore useful for cheque recognition research.

So far, for both Dari and Pashto languages, no such useful and standard handwritten databases were reported by the research communities. Hence, the two databases presented in this chapter are the first handwritten databases for Pashto and Dari languages. The beauty and usefulness of these two databases is in the selection of their contents, which cover many domains. Each database contains six types of data, i.e., isolated digits, numeral strings of various lengths, dates, isolated letters/alphabets, words and special symbols. The selected Terms/Words are those words which are usually used for the measurements of Distance, Weight and Volume, as well as words frequently used in day-to-day business activities and financial documents. These contents make our databases useful for various types of applications. In addition, we have collected handwritten full text documents for both Pashto and Dari languages. In these documents, the six types of data have been used in sentences and phrases. In the next two subsections, we will provide some background information about the Pashto and Dari languages.

3.1.1. Pashto Language

Like Arabic and Farsi languages, Pashto is a well-known language of the South-Central Region of Asia. There are more than 40 million Pashto speakers, who are mainly living in Pakistan and Afghanistan [81]. Pashto is the official language of Afghanistan as well as the provincial language of two provinces in Pakistan, i.e., North West Frontier Province (N.W.F.P) and Balochistan. In Northwest India, there are also some Pashtun communities.

The Pashto language has two main dialects, i.e., Northern Dialect and Southern Dialect. In addition, there are also some local dialects. Each dialect has some words

which are differently pronounced and/or written from the other. Hence it has been a big problem to define standard words which represent the entire Pashtun Population. To tackle with the above-mentioned challenge, we made our decision based on the fact that Pashto is the official language of Afghanistan and there exists a well-defined educational system in the Pashto language in Afghanistan. Hence, a major portion of the Data Entry Form for Pashto language has been designed based on the Afghanistan's official Pashto language. Furthermore, some equivalent words which are used in Pakistan have also been included. For example, there are two words for the English word Price, i.e., *بیټه*, which is usually used in Afghanistan and *قیمت*, which is usually used in Pakistan.

The scripts generally used for Pashto writings are called 'Naskh', which are the modified Arabic scripts [81]. In addition, Pashto language has several unique letters which do not appear in any other Arabic script [81]. Due to these facts the existing handwritten databases for Arabic and Farsi are not useful for Pashto handwritten recognition research. Hence, there was a need for creating a standard handwritten Pashto database, which could represent various features of the challenging Pashto scripts.

3.1.2. Dari Language

As mentioned earlier, the Dari language, also known as Persian, is one of the Indo-Iranian languages [82]. Dari and Pashto are the two official languages of Afghanistan, but Dari is considered to be the "lingua franca"* of Afghanistan [82]. Approximately five million people in Afghanistan and a total of two and half a million people in Iran, Pakistan, and neighboring regions, speak in the Dari language [82]. In addition, there are

* People relating to different regions and cultures who speak different languages usually communicate with each other in Dari language in Afghanistan.

hundreds of thousands of Dari speakers living as refugees and immigrants in other parts of the world, i.e., North America, Australia and Europe [82].

Dari and Farsi languages use almost the same type of modified Arabic scripts, however, in Dari the stress accent is less prominent as compared to Farsi [82]. Hence, the main difference between Dari and Farsi is in the grammar. In addition, Dari also has some loaned words from the Arabic and Farsi languages. Dari is also a cursive language similar to Arabic, Farsi and Pashto languages and is written from right to left [82]. Furthermore, Dari not only uses the Farsi numerals but also shares the same 32 letters with Farsi language. Due to this fact, our Dari database may prove to be a very useful source for Farsi handwritten recognition research.

The rest of the chapter is organized in such a way that Section 3.2 describes the data collection methodologies, whereas Section 3.3 describes the data extraction and archiving. In Section 3.4, we give a general overview of the various aspects of the two databases. In Section 3.5, the data within the two databases are described in details. Section 3.6 describes the handwritten Pashto and Dari documents. A brief discussion and future possible work are described in Section 3.7, whereas Section 3.8 concludes the chapter.

3.2. Data Collection

3.2.1. Data Entry Form

A two-page Data Entry Form, similar to that used by Solimanpour et al.[79], has been designed and used for collecting the handwritten data from Pashto and Dari speakers. Moreover, we have used an identification tag for uniquely identifying a writer

of the corresponding language. The tag consists of three English letters followed by four digits. The first three letters specify the language whereas the four digits uniquely identify the writers. For example, the Data Entry Form labeled as ‘**DAR0144**’ specifies that this form is related to the Dari language and has been written by the writer number “144”. This number is the same on both the pages of the Data Entry Form for a single writer. Similarly, in tag ‘**PSH0144**’, the three letters PSH stand for the Pashto language and 0144 specifies Pashto writer number “144”. The sample of the Data Entry Form is shown in Appendix B.

3.2.2. Writers

The process of data collection was conducted in Montreal, Canada and the Swat district of the North West Frontier Province of Pakistan. We selected North West Frontier Province in Pakistan because this province is near Afghanistan and a large number of Dari and Pashto speakers come from Afghanistan to this region for business purposes. In addition, hundreds of thousands of Afghanis, who speak Dari and/or Pashto languages, are living as refugees, immigrants and students in this province.

The writers included both males and females from various professional backgrounds, education levels and ages. We were interested in keeping track of whether the writer is male/female. In addition, we recorded whether he/she is right-handed or left-handed. Although getting this information about the writer has no significant importance at this stage, it could be used in future research work. In this way, the writers have been divided into four categories based on their genders and hand orientations as follows:

- Right-Handed Males.
- Left-Handed Males.

- Right-Handed Females.
- Left-Handed Females.

To gather this kind of information, we have put four special boxes on top of the second page of the Data Entry Form. These boxes were labeled as Male, Female, Left-handed and Right-handed and the writer had to mark the boxes that define his/her gender and hand-orientation.

Figure 3.1: Section of Form used for recording users' gender and hand-orientation.

Figure 3.1 shows the top section of one of the filled Data Entry Forms which indicates that writer number “144” in the Dari database is a Left-handed Female. Table 3.1 and Table 3.2 show the percentages of each category of writers in Pashto and Dari databases, respectively.

Table 3.1: Percentages of Pashto writers based on gender and hand-orientation.

Writers' Category	Total Number	Percentage *
Right-Handed Males	183	83.562 %
Left-Handed Males	7	3.196 %
Right-Handed Females	26	11.872 %
Left-Handed Females	3	1.369 %

Table 3.2: Percentages of Dari writers based on gender and hand-orientation.

Writers' Category	Total Number	Percentage*
Right-Handed Males	187	84.234 %
Left-Handed Males	8	3.604 %
Right-Handed Females	23	10.36 %
Left-Handed Females	4	1.802 %

* The values have been rounded off.

3.3. Data Extraction and Preparation

For any good handwritten database it is important, for all the images, to have a good quality. Achievement of this objective depends upon various factors such as the quality of scanning, the selected data extraction programs and the pre-processing methods applied at various levels.

3.3.1. Scanning

The filled forms have been scanned with 300dpi and saved as true color RGB as well as Gray Scale Images. All of the images have been stored using the Tagged Image File Format (TIFF). It uses “tags” in the file header, which enable the handling of multiple images and data in a single file. Furthermore, TIFF files have the ability to store the image data in a lossless format. Due to this ability, it is a very useful method for archiving images [83].

3.3.2. Pre-processing

We applied pre-processing at two levels, i.e., first at the form level and second at the individual image level. At the form level, those noises which were caused during scanning, such as extra noisy edges, were removed. Furthermore, the skewed or slanted images of scanned forms were corrected through computer programs. At the second level a computer program was applied to extract the data from those pre-processed forms. Once the data was extracted from the forms, the individual images were pre-processed, in order to remove the noises such as “Salt and Pepper Noise” from each extracted image. For this purpose, a median filter of kernel size 3*3 was applied, which not only removed the noise completely but also improved the quality of the images without any blurring or

distortion. The actual contents of each image were centered through computer code. The noises which had not been removed by the median filter were removed manually.

Furthermore, the extracted images were stored with 300*300 resolutions in TIFF file formats. The databases were initially created in Gray-Scale format. After that, copies of the same Gray-Scale databases have been taken and converted into Binary Image Format. Hence, we have two versions of both Pashto and Dari databases, i.e., Gray-Scale version and Binary version.

3.4. General Overview

In this section, a general overview of the various aspects of the two databases are described, i.e., directory hierarchy, naming conventions, overall statistics and the ground truth information.

3.4.1. Directory Hierarchy

The Pashto and Dari databases share the same directory hierarchy as shown in Figure 3.2. The Main Directory specifies the name of the entire database, i.e., for the Pashto language it is named as the Pashto_Database, whereas for the Dari language it is named as the Dari_Database. The arrow-heads are pointing towards the sub-directories from the root directory. The main directory further contains two sub-directories, i.e., Gray-Scale-Images and Binary-Images. Each of these two directories contains the same number of sub-directories and data but the difference is that the Gray-Scale-Images directory contains the Gray-Scale version of images of the databases while the Binary-Images directory contains the Binary version of images of the databases. Each of these two sub-directories further contains six sub-directories, which actually describe the six

types of data we have collected for corresponding language: Dates, Words, Isolated-Digits, Isolated-Characters, Special-Symbols and Numeral Strings, written in Pashto and Dari languages.

The actual data within the child-directories of each of the six sub-directories have been divided in such a way that 60% of the images have been assigned to the Training set while the Testing Set and Validation Set have each been assigned to handle 20% of images. The images inside one set, i.e., Training, are totally different from the Images of the other two sets, i.e., Testing and Validation.

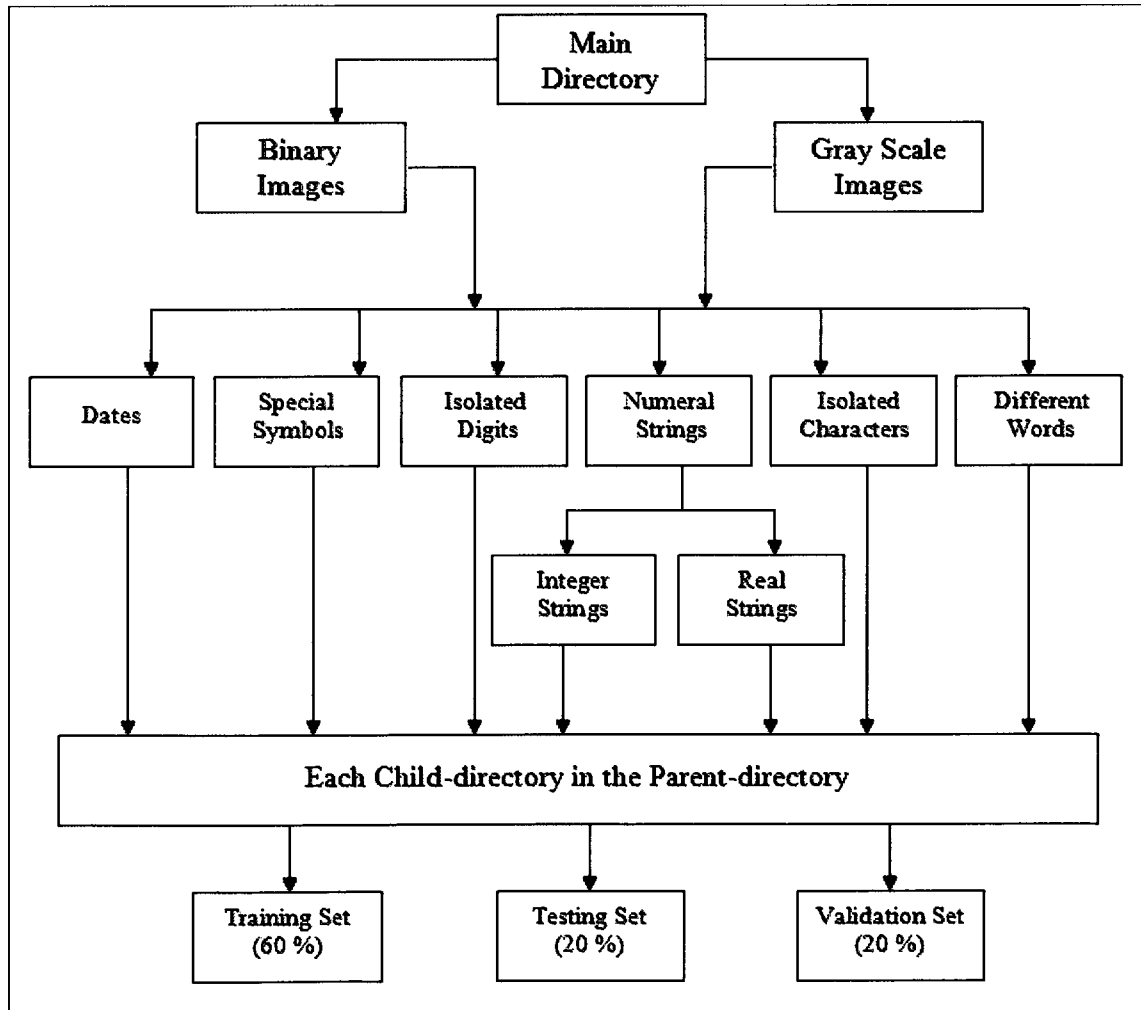


Figure 3.2: Hierarchical representation of the Databases' directories.

3.4.2. Naming Conventions for Directory

In order to make our databases more user-friendly, we have named all the folders and subfolders in such a way that the end user will be able to understand what is inside a directory by reading the name of the corresponding directory in English. For example, in both the databases, the directory which contains the different words for the corresponding language has been named as *Words*. Each sub-directory in the directory *Words* has been given the name which represents the English equivalent word of the content it contains. For example, the sub-directory '*kilometer*' inside the *Words* directory of the Pashto database contains the Pashto words which are used to represent the distance measurement unit, i.e., kilometer.

Moreover, for isolated letters from Pashto and Dari alphabets, each directory has been given the name which represents the pronunciation of the corresponding letter of each alphabet. For example, the sub-directories '*Alef*', '*Bey*', '*They*', '*Jeem*', '*Seen*' etc., contain the respective isolated letter of each alphabet.

3.4.3. Overall Statistics

As mentioned earlier, in order to make our databases ready for the experiments, we have divided the data in each directory into three disjoint sets, i.e., Training, Testing, and Validation Sets, as shown in Figure 3.2. In Table 3.3 and Table 3.4, we provide the overall statistics of Pashto and Dari databases, respectively.

The statistics shown in Table 3.3 and Table 3.4 were calculated once all the bad samples had been removed from the two databases. Due to this fact, the total number of handwritten samples for some contents is less than the total number of writers. For

example, the Pashto database contains 216 handwritten samples of Pashto dates, whereas the total number of writers was 219. On the other hand, the Dari database contains handwritings of 222 different writers, whereas the total number of handwritten samples for Dari dates was 217. For detailed statistics of both the databases, see Appendix D.

Table 3.3: Overall Statistics of the Pashto Database.

Type of Data		Classes	Total	Training Set	Testing Set	Validation Set	
Dates		1	216	128	44	44	
Isolated Characters		49	10714	6402	2156	2156	
Isolated Digits (0-9)		10	33467	20073	6697	6697	
Different Words		68	14882	8898	2992	2992	
Special Symbols		6	1313	785	264	264	
Numeral Strings	Integer Strings	Length 2	13	2892	1752	570	570
		Length 3	7	1533	919	307	307
		Length 4	6	1313	787	263	263
		Length 6	5	1095	656	219	220
		Length 7	5	1095	656	219	220
	Real Strings	Length 4	1	219	131	44	44
		Length 5	1	219	131	44	44

Table 3.4: Overall Statistics of the Dari Database.

Type of Data		Classes	Total	Training Set	Testing Set	Validation Set	
Dates		1	217	129	44	44	
Isolated Characters		37	8211	4955	1628	1628	
Isolated Digits (0-9)		10	32185	19299	6441	6445	
Different Words		73	16187	9763	3212	3212	
Special Symbols		7	1543	927	308	308	
Numeral Strings	Integer Strings	Length 2	13	2875	1725	573	577
		Length 3	7	1549	928	309	312
		Length 4	6	1300	778	260	262
		Length 6	5	1105	663	220	222
		Length 7	5	1100	658	221	221
	Real Strings	Length 4	1	222	134	44	44
		Length 5	1	216	128	44	44

3.4.4. The Ground Truth Information

Accurate and well-structured ground truth information is considered to be an essential part of a handwritten database, which makes it more useful and convenient for conducting experiments. For each type of handwritten data in our databases, we have provided a set of ground truth information which has been stored in corresponding folders for each type. Figure 3.3.a shows a snapshot of the ground truth information for an image of a numeral string in our database. For each image, the important information such as image name, writer's ID, writer's gender, writer's hand-orientation, data type, image's content, and length (i.e., number of connected components) have been provided. This ground truth information has been stored in the 'txt' file format. We have used the 'txt' file format because it is compatible to almost all platforms, e.g., Windows and Linux operating systems.


Image	
Image Name	PSH0003_P01_055.tif
Writer's ID	PSH0003
Writer's Gender	Male
Hand-Orientation	Right-Handed
Data Type	Integer String
Content (True Label)	8249001
Number of Connected Components	7

Figure 3.3.a: A sample of the ground truth information for a numeral string.

The format of ground truth information shown in Figure 3.3.a is similar for the five out of six types of data we have in our database, i.e., Words, Isolated Characters, Isolated Digits, Numeral Strings and Special Symbols. For Dates, we have included an additional field, i.e., date format because dates in both Pashto and Dari languages are written in

various formats, as will be discussed in Section 3.5.2. A snapshot of the ground truth information for Dates is shown in Figure 3.3.b.

Image	
Image Name	PSH0016_P01_001.tif
Data Type	Date
Format	yyyy/mm/dd
Content (True Label)	2007/07/06
Number of Connected Components	10
Writer's ID	PSH0016
Hand-Orientation	Right-Handed

Figure 3.3.b: The ground truth information for a handwritten sample of a date.

3.5. Data Description

The six types of collected data are Dates, Words, Isolated-Digits, Isolated-Characters, Special-Symbols and Numeral Strings written in Pashto and Dari languages.

The details of each type of data are described in this section.

3.5.1. Letters/Alphabets

Both Pashto and Dari languages use Arabic alphabets plus some additional alphabets which are unique to these languages.

3.5.1.1. Pashto Isolated Letters

Pashto has normally 40 letters in its alphabets, but three of them have further classes. For example, the letter **و** has two classes **و** and **ۆ**. Letter **ه** has 2 classes **ه** and **هـ**. Letter **ی** has 5 classes, i.e., **ی**, **ئ**, **ې**, **ې** and **ی**. When preceding the other letters in a single word, **ه** is written as **هـ**. In addition, some other letters are also used, such as **ځ** and

اے . The former, i.e., Arabic hamza-alef is used as a diacritic mark written over some other regular letters in order to produce appropriate pronunciations, i.e., ؤ, ئ and ئ. The character اے is another form of letter ي and is usually used in Pakistan, especially in the North West Frontier Province. Whereas, the letter ي is usually used in Afghanistan. The handwritten samples of all these 49 letters are shown in Figure 3.4.

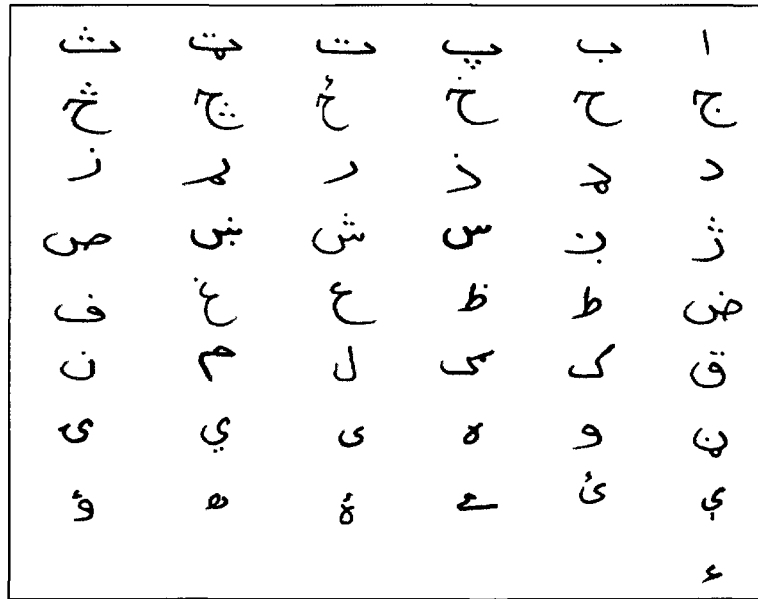


Figure 3.4: Handwritten Samples of 49 characters in the Pashto Alphabets.

3.5.1.2. Dari Isolated Letters

The Dari language is the Afghan dialect of Farsi, so it uses the same 32 letters of Farsi language. In addition to these 32 letters, we also considered five other letters, three of which are the modified forms of letter, ه, i.e., ه, ه, and ه. The letter ه represents the shape of letter ه when it is written at the initial position within a word. The other two added letters include ه (waw hamza) and the Arabic letter ه (Hamza Alef). Every writer wrote these 37 letters only once in the Data Entry Forms. Figure 3.5 shows handwritten samples of these Dari letters, taken from a filled form.

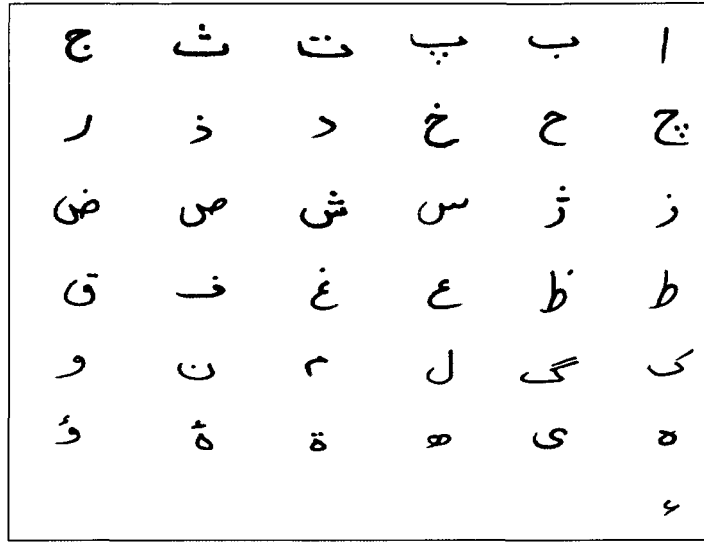


Figure 3.5: Handwritten Samples of 37 characters in the Dari Alphabets.

3.5.2. Words and Terms

A large portion of both Pashto and Dari databases consists of different words. These words include the measurement units of Distance, Volume, Weight, Currency and terms which are commonly used in financial documents and day-to-day business activities. We describe the words used in both the databases in the next sub-sections.

3.5.2.1. Words in Pashto Database

The directory in the Pashto database named as “**Words**” contains 68 sub-directories. Each sub-directory contains the handwritten samples of a different Pashto word, commonly used for describing the measurement or day-to-day business. The measurement units can be categorized into five groups. The first group consists of seven Pashto words which are commonly used for representing measurement units and distance terms. The handwritten samples of these words are shown in Figure 3.6.a.

ملي ميټر	انچه	سنټي ميټر	ميټر	كلو ميټر	اوږدوالی	پلنوالی
Millimeter	Inch	Centimeter	Meter	Kilometer	Length	Width

Figure 3.6.a: Pashto words used for expressing measurement units and distance terms.

The second group contains five Pashto words equivalent to the English words used for measurement units and terms related to Weight. Figure 3.6.b shows the handwritten samples of these words. The four Pashto words included in the third group are shown in Figure 3.6.c. These words are used for the measurement units and terms related to Volume.

ملي گرام	گرام	كلو گرام	ټن	وزن
Milligram	Gram	Kilogram	Ton	Weight

Figure 3.6.b: Pashto words used for expressing measurement units and weight terms.

ملي ليټر	ليټر	گيلون	حجم
Milliliter	Liter	Gallon	Volume

Figure 3.6.c: Pashto words used for expressing measurement units and volume terms.

The fourth group includes two Pashto words which are used to represent the names of the official currencies of Afghanistan and Pakistan, i.e., Afghani and Rupee, respectively. The handwritten samples of the two currency names are shown in Figure 3.6.d.

روپي	افغانی
Rupee	Afghani

Figure 3.6.d: Pashto words used for writing currency units.

The fifth group consists of 23 Pashto words which are used for counting quantities.

The handwritten samples of these words and their English equivalents are given in Figure 3.6.e.

شپږ	پنځه	څلور	درې	دوه	يو
Six	Five	Four	Three	Two	One
دېرش	شپږ	لس	نهه	اته	اووه
Thirty	Twenty	Ten	Nine	Eight	Seven
لوې	اتيا	اويا	شپيته	پنځوست	څلوېښت
Ninety	Eighty	Seventy	Sixty	Fifty	Forty
		کروړ	لاک	زر	سل
		10-Million	100-thousand	Thousand	Hundred

Figure 3.6.e: Pashto words for counting quantities.

The rest of the words are those Pashto words which are used in day-to-day business activities and financial documents. The handwritten samples of some of these Pashto words are shown in Figure 3.6.f.

مقدار	ټول	قرڼا	نقد	خروج	واجب الادا
Amount	Total	Credit	Cash	Cost	Due
سوار	زخيره	ټيښ	بیه	سپارل	تاوان
Interest	Stock	Tax	Price	Delivery	Decrease
درزن	پونځی	آهانہ	کرایه	گټه	ختم
Dozen	Duty	Balance	Rent	Increase	Expire
تعداد	تخير	فهرست	قسم	نېټه	
Number	Item	Inventory	Article	Period	

Figure 3.6.f: Pashto words used in day-to-day business activities and financial document.

3.5.2.2. Words in Dari Database

There are 73 different Dari words which can be found in the Dari database. Like the Pashto words, as described in the preceding section, these Dari words have also been categorized into two types, i.e., measurement units/terms and business-related words. The first type has been divided into six groups.

In the first group, we have included the three common prefixes used with various measurement units, i.e., Kilo, Centi, and Milli. These prefixes are used in combination with the basic measurement units to represent “Thousand X”, “Hundred X”, and “1/1000th X”, respectively, where X is the basic unit in question. Figure 3.7.a shows the handwritten Dari words used for representing the three mentioned prefixes.

ملي	سانتي	کيلو
Milli	Centi	Kilo

Figure 3.7.a: Prefixes of measurement units written in Dari scripts.

The second group consists of the Dari word used for the basic unit of distance, i.e., meter, and the Dari terms used for inch, length and width, as shown in Figure 3.7.b. The Dari words for Gram, Tone and Weight measurement have been included in the third group and are shown in Figure 3.7.c.

عرض	درازی	انچ	متر
Width	Length	Inch	Meter

Figure 3.7.b: Dari words used for distance measurement units and terms.

وزن	تون	گرام
Weight	Ton	Gram

Figure 3.7.c: Dari words used for weight measurement units and terms.

The fourth group consists of two Dari words used for writing the two units of Volume, i.e., Gallon and Litre. Figure 3.7.d shows the handwritten samples of these words. The fifth group contains the names of the two currencies, i.e., Afghani and Rupee, which are shown in Figure 3.7.e.

لتر	گالډن
Litre	Gallon

Figure 3.7.d: Dari words used for volume measurement units.

روپه	پول افغانی
Rupee	Afghani

Figure 3.7.e: Dari words used for writing currency units.

The 23 words of the sixth group represent some of the Dari words which are used for counting quantities. Figure 3.7.e shows the handwritten samples of these Dari words along with their English translations.

شش	پنج	چهار	سه	دو	یک
Six	Five	Four	Three	Two	One
سی	بیست	ده	نه	هشت	هفت
Thirty	Twenty	Ten	Nine	Eight	Seven
نود	هشتاد	هفتاد	شصت	پنجاه	چهل
Ninety	Eighty	Seventy	Sixty	Fifty	Forty
میلیون			لک	هزار	صد
10-Millions			100-thousands	Thousand	Hundred

Figure 3.7.f: Dari words used for counting quantity.

The second type of categorization represents those Dari words which are used in day-to-day business and financial activities, e.g., price, cost and interest, etc. Figure 3.7.g shows the handwritten images of these Dari words.

مبلغ	اعتبار	فواله کرون	قیمت	تکس	موخړی انبار
Amount	Credit	Delivery	Price	Tax	Stock
مدت	صورت کالا	ختم	فایده	نقد	میزان
Period	Inventory	Expired	Interest	Cash	Balance
ادا کړدنه	جمع	مجموع	کرایه	پارچه	جنس
Due	Plus	Total	Rent	Article	Item
انتقال	خرج	تاوان	فروش	انزارة	درجن
Transfer	Cost	Decrease	Amount	Weight	Dozen
باقي			قوٹی	زیاده	وصول
Not Received			Cotton	Increase	Received

Figure 3.7.g: Dari words and terms commonly used in day-to-day activities and financial documents.

3.5.3. Dates

The date in both Pashto and Dari languages is usually written in the format: **yyyy/mm/dd**. However, in our collected data, we found that some people have also used other formats. For example, some people have written the date in the format **dd/mm/yyyy**. In addition, some people have used the hyphen instead of the slash as the delimiter, e.g., **yyyy-mm-dd**. The handwritten samples of the three different formats found in both Pashto and Dari databases are given in Figure 3.8.a and Figure 3.8.b, respectively.

۲۰۰۷/۰۶/۲۴	۱۴/۰۸/۲۰۰۷	۲۰۰۷-۰۷-۲۱
2007/06/24	22/02/2007	2007-07-21

Figure 3.8.a: Pashto Dates written in three formats.

۲۰۰۷ ۸ / ۱۲	۲۲ / ۷ / ۲۰۰۷	۲۰۰۷ - ۰۸ - ۱۴
2007/08/12	14/08/2007	2007-08-14

Figure 3.8.b: Dari Dates written in three formats.

3.5.4. Special Symbols

We have also included some special symbols in our databases. The Pashto database contains six special symbols, two of which are currency signs that are different from currency units, i.e., *Afghani-sign* and *Rupee-sign*. The other four consist of *at-the-rate*, *number-sign*, *slash* and *colon*. It is worth mentioning that two of these symbols, i.e., *at-the-rate* and *number-sign*, are not typical Pashto symbols and people don't usually use these symbols in traditional writings. These symbols are shown in Figure 3.9.a.

@	#	/	:	؀	Rs
At-rate	Number	Slash	Colon	Afghani Sign	Rupee Sign

Figure 3.9.a: Special symbols included in the Pashto Database.

The Dari database contains seven special symbols, of which the first six are the same as those used in the Pashto database, whereas the seventh one is the *Riyal sign*. The seven symbols taken from the Dari database are shown in Figure 3.9.b.

@	#	/	:	؀	Rs	ريال
At-rate	Number	Slash	Colon	Afghani Sign	Rupee Sign	Riyal

Figure 3.9.b: Special symbols included in the Dari Database.

3.5.5. Numbers/Digits

The Pashto and Dari languages are written from right-to-left, however, the numerals in both of these languages are written from left-to-right. There are thousands of isolated digits in both the databases. The ten digits include 0 to 9 and every writer has written each digit in two formats, i.e., isolated format as well as numeral string format.

Table 3.5: Repetition frequencies of digits in Pashto and Dari Data Entry Forms.

Digit	Total Frequency	Isolated Format	In Numeral Strings
0	19	2	17
1	16	2	14
2	16	2	14
3	16	2	14
4	16	2	14
5	15	2	13
6	17	2	15
7	15	2	13
8	15	2	13
9	18	2	16

Table 3.5 shows the repetition frequencies of the Pashto and Dari digits in the Data Entry Form. In the next subsections, we will describe the Pashto and Dari numerals.

3.5.5.1. Pashto Numerals

Pashto numerals are written in Arabic style, except for the numbers 4 and 5, which are written in Farsi style. In Figure 3.10.a, the first row shows the Pashto numbers while the second row shows the corresponding English numbers.

۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
0	1	2	3	4	5	6	7	8	9

Figure 3.10.a: Handwritten Samples of Pashto Numerals.

It is important to mention that the Pashto numbers shown in Figure 3.10.a are in accordance with the numbers printed on the official currency of Afghanistan, i.e., Afghani. Most of the people write the Pashto digits in this style. However, sometimes people use ٤ (i.e., Urdu 4) instead of ٤ (i.e., Farsi 4) and ٥ (i.e., Arabic five) instead of ٥ (i.e., Farsi 5).

Moreover, in the collected data for the Pashto database, we found that there were several variations among the handwritten samples of some digits, e.g., the digit 0 has been written in two ways, i.e., either like a small filled dot or a small hollow circle. The latter style could be mistaken for the digit 5. Similarly, the digit 4 has been written in 3 ways, whereas digit 5, digit 6 and digit 7 have each been written in 2 ways. All these variations have been shown in Figure 3.10.b.

○	●	٤	٤	٤	٥	٥	٦	٧	٨	٩
0		4			5		6		7	

Figure 3.10.b: Variations found in Pashto Isolated digits database.

3.5.5.2. Dari Numeral

Like the Dari alphabets, the Dari digits are also similar to Farsi digits. Figure 3.11.a shows the handwritten samples of Dari Digits.

●	١	٢	٣	٤	٥	٦	٧	٨	٩
0	1	2	3	4	5	6	7	8	9

Figure 3.11.a: Handwritten samples of Dari Numerals.

In the Dari database, we also found several variations in the handwriting styles used for writing Dari digits by different writers. For example, for digits 0, 2, 4, 5, 6 and 7 we have seen variations as shown in Figure 3.11.b.

0	•	٢	٢	٢	٢	٢	0	0	٦	٦	٦	٧
0	2		4			5		6		7		

Figure 3.11.b: Variations found in handwritten isolated digits in Dari database.

3.5.6. Numeral Strings

Both Pashto and Dari databases contain handwritten samples of 38 classes of numeral strings. The numbers, lengths and labels of these numeral strings in the Data Entry Forms are the same for both Pashto and Dari databases. These numeral strings have various lengths from 2 to 7 digits. Out of the 38 classes of numeral strings, 36 classes consists of integer strings (i.e., they don't contain a decimal point), while the two remaining classes consist of real strings (i.e., they do contain a decimal point).

3.5.6.1. Integer Strings

The 36 classes of integer strings in each language are of five different lengths, i.e., Length 2, Length 3, Length 4, Length 6 and Length 7. There are 13 different numeral strings of 2 in each of the Pashto and Dari databases. The handwritten samples along with their English labels, taken from Pashto and Dari databases are shown in Figure 3.12.a and Figure 3.12.b, respectively.

۳۱	۷۵	۳۳	۶۹	۷۸	۴۸	۹۵
31	75	33	69	78	48	95
۲۳	۵۰	۹۹	۶۲	۶۱	۴۷	
23	50	99	62	61	47	

Figure 3.12.a: Numeral Strings of length 2, selected from the Pashto Database.

۳۱	۷۵	۳۳	۶۹	۷۸	۴۸	۹۵
31	75	33	69	78	48	95
۲۳	۵۰	۹۹	۶۲	۶۱	۴۷	
23	50	99	62	61	47	

Figure 3.12.b: Numeral Strings of length 2, selected from the Dari Database.

The seven different integer numeral strings in the Pashto and Dari databases are shown in Figure 3.13.a and Figure 3.13.b, respectively.

۹۰۸	۱۳۶	۴۲۳	۵۶۴	۰۸۶	۷۹۲	۳۹۴
908	136	423	564	086	792	394

Figure 3.13.a: Numeral Strings of length 3, selected from the Pashto Database.

۹۰۸	۱۳۶	۴۲۳	۵۶۴	۰۸۶	۷۹۲	۳۹۴
908	136	423	564	086	792	394

Figure 3.13.b: Numeral Strings of length 3, selected from the Dari Database.

There are six different integer numeral strings of length 4 in both the databases. The handwritten samples of these numeral strings for both Pashto and Dari databases are shown in Figure 3.14.a and Figure 3.14.b, respectively.

۲۲۵۶	۶۸۳۹	۷۱۰۸	۸۲۰۳	۹۷۰۱	۵۵۸۴
2256	6839	7108	8203	9701	5584

Figure 3.14.a: Numeral Strings of length 4, selected from the Pashto Database.

۲۲۵۶	۶۸۳۹	۷۱۰۸	۸۲۰۳	۹۷۰۱	۵۵۸۴
2256	6839	7108	8203	9701	5584

Figure 3.14.b: Numeral Strings of length 4, selected from the Dari Database.

The other five integer numeral strings are of length 6. The handwritten samples of these numeral strings, selected from the Pashto and Dari databases are shown in Figure 3.15.a and Figure 3.15.b, respectively.

۰۴۸۶۸۳	۲۹۳۷۱۲	۴۵۰۹۶۱	۳۶۵۲۹۷	۸۱۰۶۷۴
048283	293712	450961	365297	810674

Figure 3.15.a: Numeral Strings of length 6, selected from the Pashto Database.

۰۴۸۶۸۳	۲۹۳۷۱۲	۴۵۰۹۶۱	۳۶۵۲۹۷	۸۱۰۶۷۴
048283	293712	450961	365297	810674

Figure 3.15.b: Numeral Strings of length 6, selected from the Dari Database.

The last five classes of integer numeral strings are of length 7. Figure 3.16.a and Figure 3.16.b show the five different strings for both Pashto and Dari databases, respectively.

۸۲۴۹۰۰۱	۰۵۸۱۲۹۴	۲۴۶۰۲۵۷	۱۲۷۹۳۴۰	۱۶۷۹۴۰۰
8249001	0581294	2460257	1279340	1679400

Figure 3.16.a: Numeral Strings of length 7, selected from the Pashto Database.

۸۲۴۹۰۰۱	۰۵۸۱۲۹۴	۲۴۶۰۲۵۷	۱۲۷۹۳۴۰	۱۶۷۹۴۰۰
8249001	0581294	2460257	1279340	1679400

Figure 3.16.b: Numeral Strings of length 7, selected from the Dari Database.

3.5.6.2. Real Strings

We have also collected two classes of real strings, i.e., strings with a decimal point, in each of the databases. The decimal point in Pashto language is represented by the character ء . It is important to mention that character ء is not the *Arabic Hamza-Alef*. Rather, it represents the shortcut for *Arabic Ain* and is used as an abbreviation for the term *Ashara*, which means the tenth part of something, i.e., 1/10. Figure 3.17.a shows the Pashto handwritten real strings and their English equivalents as well as the Pashto isolated decimal point.




		
1.50	71.35	Pashto Decimal Point

Figure 3.17.a: Real Strings and decimal point written in Pashto numerals.

On the other hand, in Dari language the decimal point is represented by a slash. Figure 3.17.b shows the handwritten samples of real strings and the isolated decimal point taken from the Dari database.




		
1.50	71.35	Dari Decimal Point

Figure 3.17.b: Real Strings and decimal point written in Dari numerals.

3.6. Full Text Handwritten Documents

In addition to the two large databases described above, we have also collected two sets of handwritten documents written in Pashto and Dari languages. In these documents, the six types of data, described in the preceding sections, have been used in full real world sentences. Each writer was given some pre-printed documents and he/she had to copy the same text onto blank pages.

For the Pashto language, we prepared seven different documents which were copied by each writer. We collected a total of seventy documents written by 10 writers. For the Dari language, we prepared 5 different documents which were copied by each writer. We collected a total of 55 Dari documents written by 11 writers. Figure 3.18.a shows Pashto text taken from one of the documents. The underlined words are some of those words which have been described in Section 3.5.1. In addition, Figure 3.18.a also shows some Pashto numeral strings and a date written in Pashto format. Figure 3.18.b shows some Dari text taken from a Dari handwritten document. These documents could become a very useful source for researchers working in page segmentation and word spotting technologies in Indo-Iranian languages. Appendix C shows the samples of all the seven Pashto documents as well as the five Dari documents.

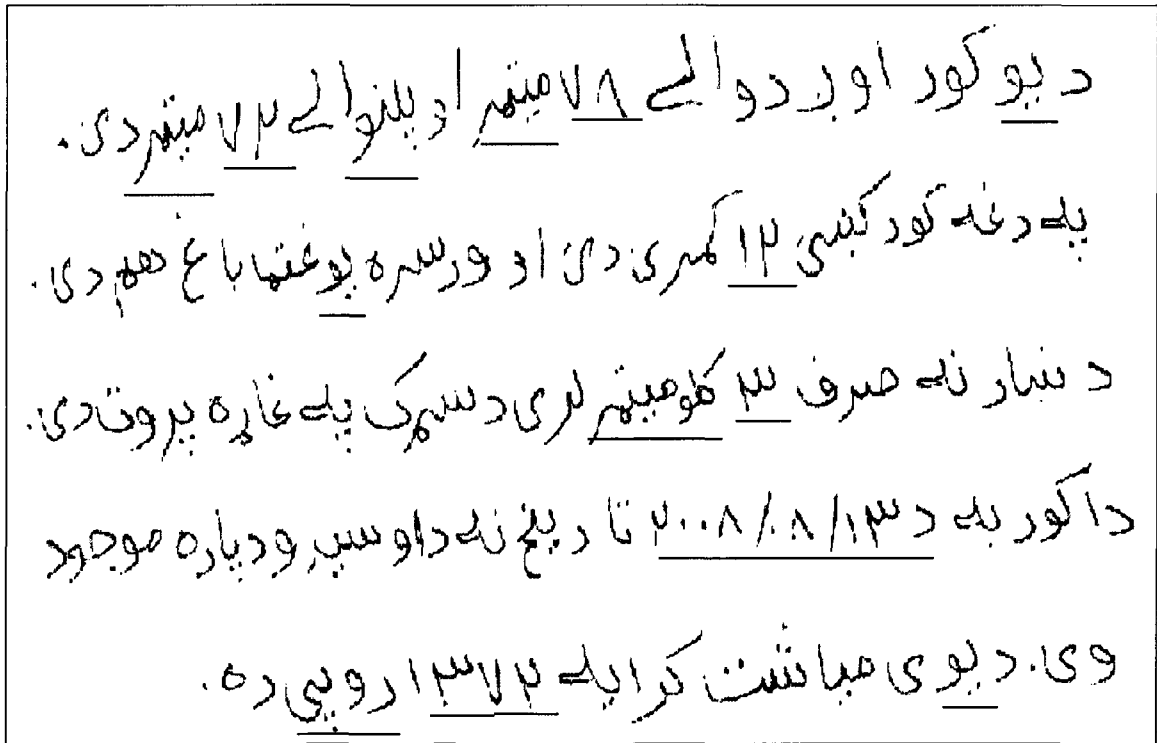


Figure 3.18.a: Part of a handwritten document written in Pashto language.

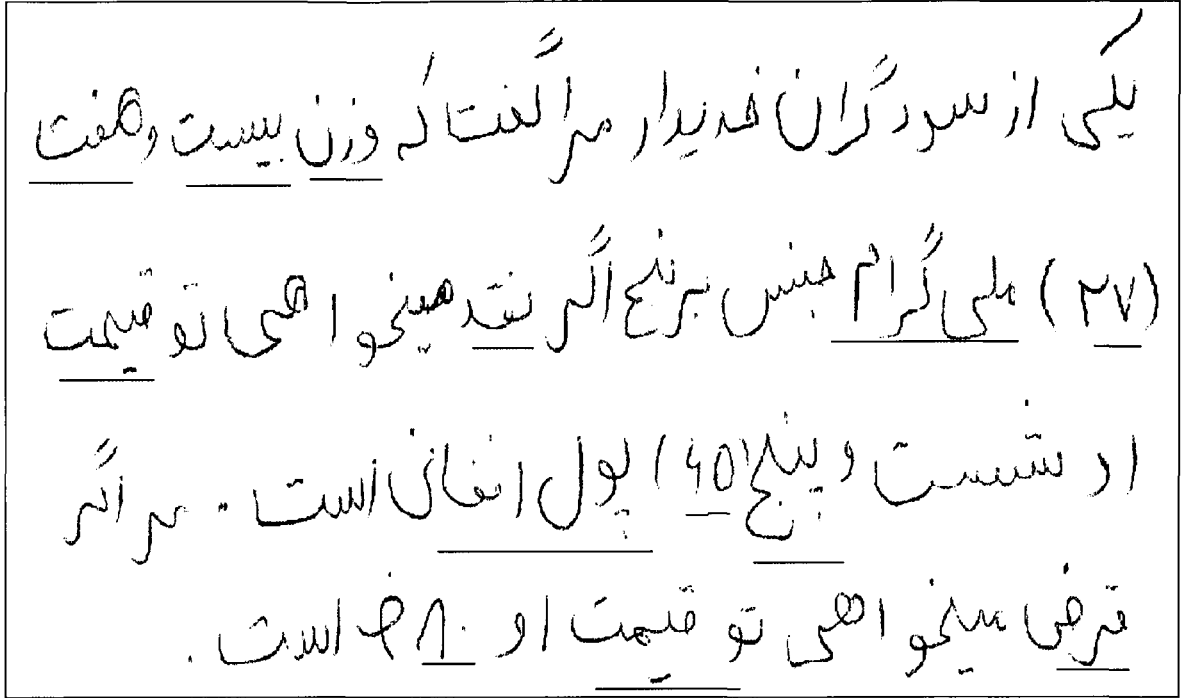


Figure 3.18.b: Part of a handwritten document written in Dari language.

To make our documents really useful for future experiments, we have also provided important ground truth information about the contents of each document. Figure 3.19 shows an example of this ground truth information. It is important to mention that such information is provided only for those contents which are related to the six types of data we have in our databases.

Document's ID	DAR_D00_W0000.tif
Content Type	Word
Number of connected components	1
Label of the content	Hundred
Top coordinates of bounding box	458
Left coordinates of bounding box	1621
Bottom coordinates of bounding box	561
Right coordinates of bounding box	1776
Height of bounding box	103
Width of bounding box	155

Figure 3.19: The ground truth information for a word within a Dari document.

3.7. Discussions and Future Work

The two databases presented in this chapter could become valuable sources of research in various domains. The contents of the two databases could become very useful in various research fields, i.e.:

- i. Isolated digit recognition
- ii. Numeral string recognition
- iii. Numeral string segmentation
- iv. Legal amount recognition
- v. Word recognition
- vi. Word segmentation
- vii. Image matching, i.e., Word Spotting
- viii. Page segmentation
- ix. Date segmentation and recognition
- x. Special symbols detection and recognition

Although these databases cover a lot of things, there is still room for further expansion and modification. The Pashto and Dari alphabets have characters with different shapes at different positions within a word. Hence, in future we could make efforts to gather handwritten data about the shapes of all these characters at the three distinct positions, i.e., first, middle, and last. More handwritten documents could be collected, which would help in various research applications, i.e., page segmentation, word spotting, indexing, information filtering and image retrieval. Furthermore, the ground truth information could be made more standardized, i.e., providing them in XML format.

3.8. Conclusion

In this chapter, the two large handwritten databases for Pashto and Dari languages have been presented. Each of these databases contains six types of data, i.e., handwritten isolated characters, isolated digits, numeral strings (different lengths), words, dates, and handwritten special symbols. These contents and the availability of ground truth information will make these databases very valuable and convenient for research experiments in various domains. Both databases are available in two formats, i.e., Gray-Scale and Binary format. The collected handwritten Pashto and Dari documents are also valuable sources for page segmentation and word spotting research. These databases will be made available in the future, for research purposes. We hope that these databases will become popular and will be useful for researchers who are interested in handwritten recognition of cursive scripts such as Pashto and Dari and other related languages, i.e., Farsi, Urdu and Arabic.

References

- [1] Rath, T. M., Manmatha, R.: Word spotting for historical documents. In: International Journal of Document Analysis and Recognition (IJDAR), Vol. 9, No. 2, pp. 139-152, (2007).
- [2] Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., Perantonis, S.J.: Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. In: International Journal of Document Analysis and Recognition (IJDAR), Vol. 9, No. 2, pp. 166-177, (2007).
- [3] Cao, H., Govindaraju, V.: Template-free word spotting in low-quality manuscripts. In: Proceedings of the 6th International Conference on Advances in Pattern Recognition (ICAPR), pp. 45-53, (2007).
- [4] Adamek, T., O'Connor, N. E., Smeaton, A. E.: Word matching using single closed contours for indexing handwritten historical documents. In: International Journal of Document Analysis and Recognition (IJDAR), Vol. 9, No. 2, pp. 153-165, (2007).
- [5] Ntzios, K., Gatos, B., Pratikakis, I., Konidaris, T.: An old Greek handwritten OCR system based on an efficient segmentation-free approach. In: International Journal of Document Analysis and Recognition (IJDAR), Vol. 9, No. 2, pp. 179-192, (2007).
- [6] Leydier, Y., Lebourgeois, F., Emptoz, H.: Text search for medieval manuscript images. In: Pattern Recognition, Vol. 40, pp. 3552-3567, (2007).
- [7] Linlin, L., Shijian, L., Chew, L. T.: A fast keyword-spotting technique. In: Proceedings of 9th International Conference on Document Analysis and

Recognition (ICDAR'07), pp. 68-72, (2007)

- [8] Srihari, S. N., Srinivasan, H., Babu, P., Bhole, C.: Spotting words in handwritten Arabic documents. In: Proceedings of SPIE, San various scripts Jose, CA, pp. 606702-1-606702-12, (2006).
- [9] Srihari, S. N., Srinivasan, H., Babu, P., Bhole, C.: Handwritten Arabic word spotting using the CEDARABIC document analysis system. In: Proceedings of Symposium on Document Image Understanding Technology (SDIUT-05), College Park, MD, pp. 123-132, (2005).
- [10] Leydier, Y., Bourgeois, F. L., Emptoz, H.: Omnilingual segmentation-free word spotting for ancient manuscripts indexation. In: Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, pp. 533-537, (2005)
- [11] Gatos, B., Konidakis, T., Ntzios, K., Pratikakis, I., Perantonis, S. J.: A segmentation-free approach for keyword search in historical typewritten documents. In: Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), pp. 54-58, (2005).
- [12] Chen, F. R., Bloomberg, D. S., Wilcox, L. D.: Spotting phrases in lines of imaged text. In: Vincent L and Baird H (Eds), Proceedings of SPIE – Document Recognition II. The International Society for Optical Engineering (SPIE) Vol. 2422, pp. 256-269, (1995).
- [13] Rath, T. M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, pp. 218-222, (2003).

- [14] Yue, L., Tan, C. L.: Word Spotting in Chinese document images without layout analysis. In: Proceedings of International Conference on Pattern Recognition (ICPR), Quebec, Canada, pp. 57-60, (2002).
- [15] Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G. V.: A line-oriented approach to word spotting in handwritten documents. In: Pattern Analysis & Applications, pp. 153–168, (2000).
- [16] Decurtins, J.: Comparison of OCR vs. Word shape recognition for keyword spotting. In: Symposium on Document Image understanding Technology, pp. 205-213, (1997).
- [17] Burl, M., Perona, P.: Using hierarchical shape models to spot keywords in cursive handwriting data. In: Proceedings of IEEE-CS Conference on Computer Vision and Pattern Recognition, pp. 535-540, (1998).
- [18] Keaton, P., Greenspan, H., Goodman, R.: Keyword spotting for cursive document retrieval. In: Proceedings of Workshop on Document Image Analysis, San Juan, Puerto Rico, pp. 74-81, (1997).
- [19] Manmatha, R., Han, C., Riseman, E. M.: Word spotting: A new approach to indexing handwriting. In: Proceedings of Computer Vision and Pattern Recognition Conference, pp. 631-637, (1996).
- [20] Kuo, S., Agazzi, O. E.: Keyword spotting in poorly printed documents using pseudo 2-d Hidden Markov Models. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 16, No. 8, pp. 842-848, (1994).
- [21] Chen, F. R., Vilcox, L. D. Bloomberg, D. S.: Word spotting in scanned images using Hidden Markov Models. In: Proceeding of the International Conference on

Acoustics, Speech and Signal Processing, Vol. 5, pp. 1-4, (1993).

- [22] Tanaka, Y., Torii, H.: Transmedia machine and its keyword search over image texts. In: Recherche d'Information assistée par Ordinateur, Cambridge, MA, pp. 248-258, (1988).
- [23] DeCurtins, J., Chen, E.: Keyword spotting via word shape recognition. In: Vincent L and Baird H (Eds.), Proceedings of the SPIE – Document Recognition II. The International Society for Optical Engineering (SPIE), Vol. 2422, pp. 270-277, (1995)
- [24] Kuo, S. S., Agazzi, O. E.: Machine vision for keyword spotting using Pseudo 2-D Hidden Markov Models. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Vol. 5, pp. 11-84, (1993).
- [25] Jeffrey, C., O'Neill, Alfred O., Hero III., Williams, W. J.: Word spotting via spatial point processes. In: Proceedings of International Conference on Image processing: IEEE, pp. 217-220, (1996).
- [26] Spitz, A. L.: Using character shape codes for word spotting in document images. In: Dori D and Bruckstein A (Eds.), Shape, Structure, and Pattern Recognition, World Scientific, Singapore, pp. 382-389, (1995).
- [27] Tan, C. L., Huang, W., Sung, S. Y., Yu, Z., Xu, Y.: Text retrieval from document images based on word shape analysis. In: Applied Intelligence, pp. 257-270, (2003).
- [28] Chen, F. R., Wilcox, L., Bloomberg, D.: Detecting and locating partially specified keywords in scanned images using Hidden Markov Models. In: Proceedings of the 2nd International Conference on Document Analysis and Recognition. IEEE

- Computer Society Press, pp. 133-138, (1993).
- [29] Rath, T. M., Manmatha, R.: Word image matching using Dynamic Time Warping. In: Proceedings of IEEE Computer Society on Computer Vision and Pattern Recognition, Vol. 2, pp. 521-527, (2003).
- [30] Belongie, S., Malik, J.: Matching with shape contexts. In: Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 20-26, (2000).
- [31] Tubbs, J. D.: A note on binary template matching. In: Pattern Recognition, Vol. 22, No. 4, pp. 359-356, (19989).
- [32] Manmatha, R., Rath, T. M.: Indexing of handwritten historical documents—recent progress. In: Proceedings of Symposium on Document Image Understanding Technology (SDIUT), pp. 77-85, (2003).
- [33] Jain, A. K., Namboodiri A. M.: Indexing and retrieval of on-line handwritten documents. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), Vol. 2, pp. 655-659, (2003).
- [34] Al-Khatib, W. G., Shahab, S. A., Mahmoud, S. A.: Digital library framework for Arabic manuscripts. In: Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, pp. 458-465, (2007).
- [35] Rothfeder, J. L., Feng, S., Rath, T. M.: Using corner feature correspondences to rank word images by similarity. In: Proceedings of Conference on Computer Vision and Pattern Recognition Workshop, pp. 30-35, (2003).
- [36] Manmatha, R., Han, C., Riseman, E. M., Croft, W. B.: Indexing handwriting using word matching. In: Proceedings of Digital Libraries'96, 1st ACM International

- Conference on Digital Libraries, pp. 151–159, (1996).
- [37] Shijian, L., Tan, C. L.: Keyword spotting and retrieval of document images captured by a digital camera. In: Proceedings of International Conference on Document Analysis and Recognition (ICDAR'07), pp. 23-26, (2007).
- [38] Shahab, S. A., Al-khatib, W. G., Mahmoud, S. A.: Computer aided indexing of historical manuscripts. In Proceedings of International Conference on Computer Graphics, Imaging and Vision. IEEE Computer Society, pp. 287-295, (2006).
- [39] Tomai, C. I., Zhang, B., Govindaraju, V.: Transcript mapping for historic handwritten document images. In: Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), pp. 413–418, (2002).
- [40] Khoubyari, S., Hull, J. J.: Keyword location in noisy document image. In: Proceedings of Second Annual Symposium on Document Analysis and Information Retrieval, UNLV, Las Vegas, pp. 217-231, (1993).
- [41] Zhang, B., Srihari, S. N., Huang, C.: Word image retrieval using binary features. In: Document Recognition and Retrieval XI, SPIE Vol. 5296, pp. 45-53, (2004).
- [42] Yue, L., Tan, C. L.: Chinese word searching in imaged documents. In: International Journal of Pattern Recognition and Artificial Intelligence, Vol. 18, No. 2, pp. 229-246, (2004).
- [43] Yue, L., Tan, C. L.: Information Retrieval in document image databases. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, pp. 1398-1410, (2004).
- [44] Yue, L., Tan, C. L.: Keyword searching in compressed document images. In: Proceedings of Data Compression Conference, pp. 25-27, (2003).

- [45] Yue, L., Tan, C. L.: Word searching in document images using word portion matching. In: Proceedings of 5th IAPR International Workshop on Document Analysis Systems, pp. 19-21, (2002).
- [46] Srihari, S. N., Srinivasan, H., Huang, C., Shetty, S.: Spotting words in Latin, Devanagari and Arabic scripts. In: Indian Journal of Artificial Intelligence, Vol. 16, No. 3, pp. 2-9, (2006).
- [47] Manmatha, R.: Multimedia indexing and retrieval research at the Center for Intelligent Information Retrieval. In: Proceedings of Symposium on Document Image Understanding Technology, pp.16-30, (1997).
- [48] Cho, B-J., Kim, J. H.: Print keyword spotting with dynamically synthesized pseudo 2D HMMs. In: Pattern Recognition Letters, Vol. 25, pp. 999-1011, (2004).
- [49] Soo, H. K., Sang, C. P., Chang, B. J., Ji, S. K., Park, H. R., Guee, S. L.: Keyword spotting on Korean document images by matching the keyword image. In: Digital Libraries: Implementing Strategies and Sharing Experiences, Vol. 3815, pp. 158-166, (2005).
- [50] Park, S. C., Son, H. J., Jeong, C. B., Soo, H. K.: Keyword spotting on Hangul document images using two-level image-to-image matching. In: Proceedings of the 18th International Conference on Innovations in Applied Artificial Intelligence, pp. 79-81, (2005).
- [51] Jawahar, C. V., Balasubramanian, A., Meshesha, M.: Word-level access to document image datasets. In: Proceedings of the Workshop on Computer Vision, Graphics and Image Processing (WCVGIP), pp.73-76, (2004).
- [52] Balasubramanian, A., Meshesha, M., Jawahar, C. V.: Retrieval from document

- image collections. In: Proceedings of Seventh IAPR Workshop on Document Analysis Systems, (LNCS 3872), pp. 1-12, (2006).
- [53] Christophe, C.: Dynamic handwritten keyword spotting based on the NSHP-HMM. In: Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR'07), pp. 242-246, (2007).
- [54] Josep, L., Gemma, S.: Indexing historical documents by word shape signatures. In: 9th International Conference on Document Analysis and Recognition (ICDAR'07), Vol. 1, pp. 362-366, (2007).
- [55] Rohlicek, J. R., Russell, W., Roukos, S., Gish, H.: Continuous Hidden Markov Modeling for speaker-independent word spotting. In: Proceedings of IEEE ICASSP, pp. 627-630, (1989).
- [56] Paul D. B., Rose, R. C.: A Hidden Markov Model based keyword recognition system. In: Proceedings of IEEE ICASSP, pp. 129-132, (1990).
- [57] Doermann, D.: The Indexing and retrieval of document images: a survey. *Computer Vision and Image Understanding: CVIU*, Vol. 70, No. 3, pp. 287-298, (1998).
- [58] Mitra, M., Chaudhuri, B. B.: Information retrieval from documents : a survey. In: *Information Retrieval*, Vol. 2, pp. 141-163, (2000).
- [59] Laurence, L-S., Abderrazak Z., Bruno T.: Text line segmentation of historical documents: a survey. In: *International Journal of Document Analysis and Recognition (IJ DAR'07)*, pp. 123-138, (2007).
- [60] Mahadevan, U., Nagabushnam, R. C.: Gap metrics for word separation in handwritten lines. In: *Proceedings of the Third International Conference on*

- Document Analysis and Recognition, Los Alamitos, CA, pp. 124-127, (1995).
- [61] Favata, J. T., Srikantan, G.: A multiple feature/resolution approach to hand-printed digit and character recognition. In: International Journal of Imaging Systems and Technology, Vol. 7, pp. 304-311, (1996).
- [62] Scott, G. L., Longuet-Higgins, H. C.: An Algorithm for associating the features of two patterns. In: Proceedings of the Royal Society of London, B224, pp. 21-26, (1991).
- [63] Manmatha, R., Rothfeder, J.: A scale space approach for automatically segmenting words from historical handwritten documents. In: IEEE Transactions, Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1212-1225, (2005).
- [64] Manmatha, R., Srimal, N.: Scale space technique for word segmentation in handwritten manuscripts. In: Proceedings of 2nd International Conference on Scale-Space Theories in Computer Vision, pp. 22–33, (1999).
- [65] Doermann D.: The indexing and retrieval of document images: a survey. In: Computer Vision and Image Understanding, Vol. 70, No. 3, pp. 287-298, (1998).
- [66] Rath, T. M., Lavrenko, V., Manmatha, R.: A search engine for historical manuscript images. In: Proceedings of the 27th Annual International ACM SIGIR Conference Sheffield, pp. 369-376, (2004).
- [67] <http://www.wikipedia.org>. Current: February 12, 2009.
- [68] Bokser, M.: Omnidocument technologies. In: Proceedings of IEEE, Vol. 80, No. 7, pp.1066-1078, (1992).

- [69] Agrawal, M., Kumar, P. M. N. S. S. K., Jawahar, C. V.: Indexing and retrieval of devanagari text from printed documents. In: Proceedings of NCDAR, pp. 244-251, (2003).
- [70] Suen, C. Y., Xu, Q., Lam, L.: Automatic recognition of handwritten data on cheques –fact or fiction? In: Pattern Recognition Letters, Vol. 20, pp. 1287-1295, (1999).
- [71] Fan, R., Lam, L., Suen, C. Y.: Processing of date information on cheques. In: Progress in Handwriting Recognition, World Scientific, pp. 473-479, (1997).
- [72] Wang, C., Hotta, Y., Suwa, M., Naoi, Satoshi.: Handwritten Chinese address recognition. In: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), pp. 539-544, (2004).
- [73] Kabir, E., Downton, A. C., Birch, R.: Recognition and verification of postcodes in handwritten and hand-printed addresses. In: Proceedings of 10th International Conference on Pattern Recognition, Vol. 1, pp. 469-473, (1990).
- [74] Lee, C. K., Leedham, C. G.: A new hybrid approach to handwritten address verification. In: International Journal of Computer Vision, Vol. 57, No. 2, pp. 107-120, (2004).
- [75] Precision and recall: http://en.wikipedia.org/wiki/Precision_and_recall. Current: February 12, 2009.
- [76] trec_eval : <http://www.emse.fr/~mbeig/IR/tools.html>. Current: February 12, 2009.
- [77] Scripts by Name: <http://www.ontopia.net/i18n/scripts.jsp>. Current: February 12,

2009.

- [78] Zhang, B., Srihari, S. N.: Binary Vector Dissimilarity Measures for Handwriting Identification. In: Proceedings SPIE, Document Recognition and Retrieval X, pp. 155-166, (2003).
- [79] Solimanpour, F., Sadri, J., Suen, C. Y.: Standard Databases for Recognition of Handwritten Digits, Numerical Strings, Legal Amounts, Letters and Dates in Farsi Language. In: Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR), pp. 3-7, (2006).
- [80] Al-Ohali, Y., Cheriet, M., Suen, C. Y.: Databases for recognition of handwritten Arabic cheques. In: Proceedings of the Seventh Int. Workshop on Frontiers in Handwritten Recognition, pp. 601-606, (2000).
- [81] Pashto Language: http://en.wikipedia.org/wiki/Pashto_language. Current: February 12, 2009.
- [82] Dari Language: http://en.wikipedia.org/wiki/Dari_language. Current: February 12, 2009.
- [83] Image File Format: http://en.wikipedia.org/wiki/Tagged_Image_File_Format. Current: February 12, 2009.
- [84] Pratt, W. K.: *Digital Image Processing*, Wiley, New York, 1978.
- [85] Saabni, R., El-Sana, J.: Keyword Searching for Arabic Handwritten Documents. In: Proceedings of 11th International Conference on Frontiers in Handwritten Recognition, pp. 271-277, (2008).
- [86] Favata, J. T., Srikantan, G., Srihari, S. N.: Handprinted Character/Digit Recognition using a Multiple Feature/Resolution Philosophy. In: Proceedings of

the 4th International Workshop on the Frontiers of Handwriting Recognition (IWFHR'94), pp. 57-66, (1994).

- [87] Zhang, B., Srihari, S. N., Lee, S.: Individuality of Handwritten Characters. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), pp. 1086-1090, (2003).
- [88] Srihari, S. N., Cha, S.-H., Arora, H., Lee, S.: Individuality of handwriting. In: Journal of Forensic Sciences, Vol. 47, pp. 1-17, (2002).
- [89] Zhang, B., Srihari, S. N.: Analysis of Handwriting Individuality Using Word Features. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), pp. 1142-1146, (2003).

Related Publications

1. Shah, M. I., Suen, C. Y.: Word Spotting Techniques in Document Analysis and Retrieval – A Comprehensive Survey. Accepted for: Handbook of Pattern Recognition and Computer Vision, 4th Edition, Edited by Prof. C.H. Chen, to be published by World Scientific Publishing, January 2010.
2. Shah, M. I., Sadri, J., Suen, C. Y., Nobile, N.: A New Multipurpose Comprehensive Database for Handwritten Dari Recognition. In: Proceedings of the 11th International Conference on Frontiers in Handwritten Recognition (ICFHR'2008), Montreal, Canada, pp. 635-641, August 19-21, 2008.
3. Shah, M. I., He, C. L., Nobile, N., Suen, C. Y.: A Handwritten Pashto Database with Multi-Aspects for Handwritten Recognition. Submitted to: 14th Conference of the International Graphonomics Society (IGS'2009), Dijon, France, September 13-16, 2009.

Appendix A:

Table A.1: Binary Equivalents of decimal values ranging from 0 to 255

Decimal	Binary	Decimal	Binary	Decimal	Binary	Decimal	Binary
0	00000000	64	01000000	128	10000000	192	11000000
1	00000001	65	01000001	129	10000001	193	11000001
2	00000010	66	01000010	130	10000010	194	11000010
3	00000011	67	01000011	131	10000011	195	11000011
4	00000100	68	01000100	132	10000100	196	11000100
5	00000101	69	01000101	133	10000101	197	11000101
6	00000110	70	01000110	134	10000110	198	11000110
7	00000111	71	01000111	135	10000111	199	11000111
8	00001000	72	01001000	136	10001000	200	11001000
9	00001001	73	01001001	137	10001001	201	11001001
10	00001010	74	01001010	138	10001010	202	11001010
11	00001011	75	01001011	139	10001011	203	11001011
12	00001100	76	01001100	140	10001100	204	11001100
13	00001101	77	01001101	141	10001101	205	11001101
14	00001110	78	01001110	142	10001110	206	11001110
15	00001111	79	01001111	143	10001111	207	11001111
16	00010000	80	01010000	144	10010000	208	11010000
17	00010001	81	01010001	145	10010001	209	11010001
18	00010010	82	01010010	146	10010010	210	11010010
19	00010011	83	01010011	147	10010011	211	11010011
20	00010100	84	01010100	148	10010100	212	11010100
21	00010101	85	01010101	149	10010101	213	11010101
22	00010110	86	01010110	150	10010110	214	11010110
23	00010111	87	01010111	151	10010111	215	11010111
24	00011000	88	01011000	152	10011000	216	11011000
25	00011001	89	01011001	153	10011001	217	11011001
26	00011010	90	01011010	154	10011010	218	11011010
27	00011011	91	01011011	155	10011011	219	11011011
28	00011100	92	01011100	156	10011100	220	11011100
29	00011101	93	01011101	157	10011101	221	11011101
30	00011110	94	01011110	158	10011110	222	11011110
31	00011111	95	01011111	159	10011111	223	11011111
32	00100000	96	01100000	160	10100000	224	11100000
33	00100001	97	01100001	161	10100001	225	11100001
34	00100010	98	01100010	162	10100010	226	11100010
35	00100011	99	01100011	163	10100011	227	11100011
36	00100100	100	01100100	164	10100100	228	11100100
37	00100101	101	01100101	165	10100101	229	11100101
38	00100110	102	01100110	166	10100110	230	11100110

39	00100111	103	01100111	167	10100111	231	11100111
40	00101000	104	01101000	168	10101000	232	11101000
41	00101001	105	01101001	169	10101001	233	11101001
42	00101010	106	01101010	170	10101010	234	11101010
43	00101011	107	01101011	171	10101011	235	11101011
44	00101100	108	01101100	172	10101100	236	11101100
45	00101101	109	01101101	173	10101101	237	11101101
46	00101110	110	01101110	174	10101110	238	11101110
47	00101111	111	01101111	175	10101111	239	11101111
48	00110000	112	01110000	176	10110000	240	11110000
49	00110001	113	01110001	177	10110001	241	11110001
50	00110010	114	01110010	178	10110010	242	11110010
51	00110011	115	01110011	179	10110011	243	11110011
52	00110100	116	01110100	180	10110100	244	11110100
53	00110101	117	01110101	181	10110101	245	11110101
54	00110110	118	01110110	182	10110110	246	11110110
55	00110111	119	01110111	183	10110111	247	11110111
56	00111000	120	01111000	184	10111000	248	11111000
57	00111001	121	01111001	185	10111001	249	11111001
58	00111010	122	01111010	186	10111010	250	11111010
59	00111011	123	01111011	187	10111011	251	11111011
60	00111100	124	01111100	188	10111100	252	11111100
61	00111101	125	01111101	189	10111101	253	11111101
62	00111110	126	01111110	190	10111110	254	11111110
63	00111111	127	01111111	191	10111111	255	11111111

(Source: http://www.dewassoc.com/support/msdos/decimal_hexadecimal.htm)

Appendix B: Sample of Data Entry Form

R
DAR0144
M

<p style="font-size: 0.8em;">در اینجا نوشته نکنید</p>	<p>Dari Handwritten Collection Form Concordia University (Montreal, Canada) Email: mu_shah@encs.concordia.ca</p>	<p>Address: 1455 de Maisonneuve W - EV3.403, Montreal QC H3G 1M8, Canada http://www.cenparmi.concordia.ca</p>
<p>میان از همکاری شما زیاده شکر گزار هستیم. این در زبان دری یک خدمت بزرگ است. ازین پراجکت اصلی مقصدش این است که درلسان دری خطبای که بدمستبا نوشته شده باشد آنرا بذریعہ کمپیوتر خوانده شود. لطفاً مهربانی کنید که خانه های زیرین را همراہ احتیاط بقلم سیاہ پُر کنید.</p>		
<p>تو کمائی کرده میتوانی عدد یک را</p>		<p>ای میل:</p>

۲	۶	۹	۱	۴	۰	۷	۵	۸	۳	۹	۸	۷	۶	۵	۴	۳	۲	۱	.	مؤرخه	
۲	۶	۹	۱	۴	-	۷	۵	۸	۳	۹	۸	۷	۶	۵	۴	۳	۲	۱	-	۰۷/۸/۱۴	
۶۳	۵۰	۷۸	۹۹	۶۱	۴۷	۶۹	۹۵	۳۳	۴۸	۶۲	۷۵	۳۱									
۴۳	۵۰	۷۸	۹۹	۴۱	۴۷	۴۹	۹۵	۳۳	۴۸	۴۲	۷۵	۳۱									
۰۸۶	۳۹۴	۱۳۶	۷۹۲	۵۶۴	۹۰۸	۴۲۳	۱/۵۰														
-۸۴	۳۹۴	۱۳۴	۷۹۲	۵۶۴	۹۰۸	۴۲۳	۱/۵۰														
۷۱۰۸	۲۲۵۶	۹۷۰۱	۵۵۸۴	۶۸۳۹	۸۲۰۳	۷۱/۳۵															
۷۱۰۸	۲۲۵۴	۹۷۰۱	۵۵۸۴	۶۸۳۹	۸۲۰۳	۷۱/۳۵															
۸۱۰۶۷۴	۳۶۵۲۹۷	۴۵۰۹۶۱	۲۹۳۷۱۲	۰۴۸۶۸۳																	
۸۱-۶۷۴	۳۶۵۲۹۷	۴۵۰۹۶۱	۲۹۳۷۱۲	-۴۸۶۸۳																	
۱۶۷۹۴۰۰	۱۲۷۹۳۴۰	۲۴۶۰۲۵۷	۰۵۸۱۲۹۴	۸۲۴۹۰۰۱																	
۱۴۷۹۴۰۰	۱۲۷۹۳۴۰	۲۴۶۰۲۵۷	-۵۸۱۲۹۴	۸۲۴۹۰۰۱																	

ذ	د	خ	ح	چ	ج	ث	ت	پ	ب	ا
ذ	د	خ	ح	چ	ج	ث	ت	پ	ب	ا
غ	ع	ظ	ط	ض	ص	ش	س	ژ	ز	ر
غ	ع	ظ	ط	ض	ص	ش	س	ژ	ز	ر
ء	ه	و	و	ن	م	ل	گ	ک	ق	ف
ء	ه	و	و	ن	م	ل	گ	ک	ق	ف
بیال	ف	Rs.	/	:	@	#	ف	ه	ة	ی
ریال	ف	Rs.	/	:	@	#	و	ه	ة	ی

Figure B.1: The first page of Data Entry Form used for data collection

DAR0144

<input type="checkbox"/> <input type="checkbox"/> در اینجا نشانه نکنید	<input type="checkbox"/> Male: مرد <input checked="" type="checkbox"/> Female: زن	<input checked="" type="checkbox"/> Left-Handed: چپ دست <input type="checkbox"/> Right-Handed: راست دست	عمر _____
---	--	--	-----------

هفت	شش	پنج	چهار	سه	دو	یک
هفت	شش	پنج	چهار	سه	دو	یک
پنجاه	چهل	سی	بیست	ده	نه	هشت
پنجاه	چهل	سی	بیست	ده	نه	هشت
یک	هزار	صد	نود	هشتاد	هفتاد	شصت
یک	هزار	صد	نود	هشتاد	هفتاد	شصت
حصول	تاوان	فایده	تکس	میزان	نقد	ملیون
حصول	تاوان	فایده	تکس	میزان	نقد	ملیون
حواله کردن	قلم	مقدار	تعداد	مبلغ	مجموع	باقی
حواله کردن	قلم	مقدار	تعداد	مبلغ	مجموع	باقی
سپردن	ختم	کرایه	وزن	اعتبار	نوع	مدت
سپردن	ختم	کرایه	وزن	اعتبار	نوع	مدت
انچ	قرض	قیمت	خرج	درازی	عرض	اندازه
انچ	قرض	قیمت	خرج	درازی	عرض	اندازه
پرداخت	جنس	زیاده	پیداوار	بدهی	فروش	پارچه
پرداخت	جنس	زیاده	پیداوار	بدهی	فروش	پارچه
کیلو	تن	گیلن	جمع	انتقال	قوطني	درجن
کیلو	تن	گیلن	جمع	انتقال	قوطني	درجن
صورت کالا	صورت کمال	متر	لتر	گرام	سانتی	ملي
صورت کالا	صورت کمال	متر	لتر	گرام	سانتی	ملي
موجودی انبار	موجودی انبار	ادا کردنش	ادا کردنش	پول افغانی	پول افغانی	پيسه
موجودی انبار	موجودی انبار	ادا کردنش	ادا کردنش	پول افغانی	پول افغانی	پيسه

Figure B.2: The second page of Data Entry Form used for data collection

Appendix C: Pashto and Dari Handwritten Documents

PSH_Dol_W0008

اقرار نامه

تاریخ: ۲۰۱۱/۰۹/۲۰

خه شير خان د کيل د اقرار کوم چه مادکا بنجو د جاویر
نه ۹۲۸۷۲ روپی قرض واغستی. دارقم به خه د ۲۰۱۱/۰۳/۱۷
نه مبنکی شير خان ته واپس کوم.
کچری د ایتکه تیره نشوه او مادارقم ادا نه کونوبیا به خه به
واجب الادا رقم باندي په زر پسي شل روپی سود هم ادا کوم.

ولیکلی شو چه سند شتی.

کواه نمبر ۲

کواه نمبر ۱

اسماعیل د کا بنجو

سلیم د شاه د پرهی

Figure C.1: Sample of handwritten document # 1 written in Pashto scripts

PSH-Doz-W0008

د يو کور او بډواله ۸ ۷ ميټر او پلنواله ۷ ۲ ميټر دی .
په دغه کور کې ۱۲ کمرې دي او ورسره يو غنما باغ هم دی .
د نبار نه صرف ۳ کلو ميټر لرې د سړک په غاړه پروت دی .
دا کور به د ۲۰۰۸/۸/۱۳ تا د پنځ زک د او سپر و د پارو موجود
وي . د يوې مياشتت کرایه ۱۳۷۲ روپي ده .

خواهشمند حضرت په دې پته را لټه کولی شئ .

سليمان شاه کابنجو

ټيليفون: ۸۱۲۸۱۹

Figure C.2: Sample of handwritten document # 2 written in Pashto scripts

PSH-D03-W0008

پرون په سوات کښې ۳۱۲ ملی میټر باران او شپږ ډیسیندو سطح
د خطر د حرنه نهه ستنی میټر لوړه شوه.

سیلاب په څیر مقدار سره فلوونډه د منځه یوړل او شپږ ډزل
تفرې هم لاسو کول. د هر قسېم څینر بیه څیره لوړه شوه.
زمیندارانو د حکومت نه د مالی مرستی او د زرعی ټیکس د
صافی غوښتنه او کړه.

په راټولو نورو دري ورځو کښې د لوړ باران هم امکان دي.

Figure C.3: Sample of handwritten document # 3 written in Pashto scripts

PSH-Do4-W0008

د کال ۲۰۰۲ نه منځنۍ د افغانۍ قدره ډیر کم وو. په یوه روپۍ
په ۱۷۳۲ افغانۍ کېدلی.
په کال ۲۰۰۸ کېنۍ د افغانۍ قدره د روپۍ په نسبت ډیر لوړدی.
نن یوه افغانۍ په څلور روپۍ خرڅېښی.
په پاکستان کېنۍ په سبزبانو بانډی چو نځنی نور زیات شو.
د سره دکاډو کرایه هم ډیره لوړه شوه.
زمینداران ټول د تاوان سره مخ دی. د تیرۍ پوی هفتی نه بازار
ته د سبزبانو سپارل بند شوی دی.

Figure C.4: Sample of handwritten document # 4 written in Pashto scripts

قدر دانه سلیم روره

خما سره په فهرست کښې چې یو نور څیزونه هم کم دی خو فصلتال
دلا ندینو چې یو ضرورت دی.

۱: د غوړولس چې بی چه وزن یې پنځه کلوگرام وی.

۲: د تیلو دوه درامونه چه حجم یې څلور کیلن او اته لیتر وی.

۳: یو تن غنم.

۴: یو انچ پتی چه تعداد یې سل وی.

مهر بانې او کړه ماته دا څیزونه راواسته و. رقم څه پد کھاته
کښې ولیکه.

شکر گزار

بخت خان د کبل

تاریخ: ۲۰۰۸/۰۹/۱۳

PSH-D06-W0008

د کال ۲۰۰۸ په اول کښی په ذخیره کښی موجود چیزونو ټول

تعداد لکه زره وو.

د ۲۰۰۸/۹/۳۰ پوری اته زده چیزونه خرڅ شول. په دی

کښی شپږ زره سل چیزونه په نقد او باقی په قرض لامل

په نقد خرڅ شوي چیزونو کښی پنځه زره اووه سوه چیزونو گڼه

او کړه او نورو تاوان او کړو.

بل طرفته په قرض خرڅ شوي ټولو چیزونو تاوان او کړو.

Figure C.6: Sample of handwritten document # 6 written in Pashto scripts

PSH-Do7-W0008

نرخنامه

تاریخ : ۲۰۰۸/۰۹/۱۵

قیمت	مقدار	خیز
۲۵۴۰۲ ری یا RS . ۲۲۰۲۳	۳۱۲۳ گرام	۱: سره
۱۶۳۲۶ ری یا RS . ۸۹۳۲	۱۸ ملی لیتر	۲: پاره
۳۰۲ ری یا RS. ۷۴۳	۱۰۰ ملی گرام	۳: زعفران

پته: کبیر کریانک ستور سینکوره سوانا

تیلیفون #: ۲۹۸۱۲۳۲

SHah @ swat.net

Figure C.7: Sample of handwritten document # 7 written in Pashto scripts

من مشتاق هادی پسر فضل هادی سهرجری
اقرار می‌کنم که از خرید شیرماهر گیلن
تیل که وزن هیندار لیتر است. بنام بیخ
۲۰۰۸/۰۹/۱۳ و پول کردم.

قیمت این مهر گیلن تیل ده هزار نه
صد و هشت پول افغانی است.

این قیمت بر ما اعتبار باقی است.
نوشتۀ ام که یاد باشد در مستقبل

تاریخ: ۲۰۰۸/۰۹/۱۸

Ao

Figure C.8: Sample of handwritten document # 1 written in Dari scripts

یکے از سودگران خریدار مراگفت که وزن
بسیت و هفت (۲۷) ملی گرام جنس برنج
اگر نقدی خواهی تو قیمت او شصت و پنج
(۶۵) پول افغانی است و اگر قرض فی
خواهی تو قیمت او ۸۰۰۰۰ است.
چه فرمائید علمائے دین درین مسئلہ
جائز است یا نیست ؟

یک شخص در ادا کردن کرایه خانۀ خود سخت
 تجلیل بود . مالک خانۀ او اورا بسیار مدت
 صبر کرد . لیکن هاشم شخص کرایه را حواله نکرد
 مالک آنخانۀ او را در ذکوة ۵۰ هزار و شش
 پول داد . و باز مطالبه کرد که از این رقم کرایه داید
 تجلیل ۵۰ هزار و دو پول جمع کرد . و اوانداد
 و مصارف خود این رقم را تفریح کرد .
 و قائده این رقم خود بخورد .
 تاریخ نوشتن این قصه : ۱۳/۳/۲۰۰۶

من حضور هادی پس از ثبت مندرجه ضروری
 اسرار محمد شاه اجناس مذکور زایلین فروخت می کنم

ادکردنش قیمت این اجناس نقد و وصول می کنم

جنس	مبلغ	قیمت	تکس	مجموع
۱: گندم	سه تن	۶۰۰۰	۱۵	۴۰۱۵
۲: صنج	چهار کیلو	۴۰۰	۲	۱۷۰۰

۷۷۱۵ میزان

مهرت کمال جمع کردم در صورتی انبار وصول کردم
 و قیمت این اجناس قبیل اصدت حواله کردن
 وصول کردم -

تاریخ : ۱۳۶/۱۲/۰۵

Figure C.11: Sample of handwritten document # 4 written in Dari scripts

نوخ ریال وپس پاکستانی بمقابلہ پول افغانی

درجہ ذیل است

۱۰۰ ریال = ۱۵۰ پ

۱۰۰ Rs = ۹۵ پ

نوخ لباس نارینه درجہ ذیل است

درازی عرض قیمت

۵ صتره یک صتره دو اینج ۲۰۰ پول

۳ صتره یک صتره ۱۵۰ پول

برائے رابطہ : خدام آغا صراف کابل

تلفون # : ۳۱۵۵۴۸۲۱

ای میل : mansoor@yahoo.com

Figure C.12: Sample of handwritten document # 5 written in Dari scripts

Appendix D: Detailed Statistics of the Dari and Pashto Databases

Table D.1: Detailed Statistics of Dari Database

Type of Data		Total	Training	Testing	Validation
Dates		217	129	44	44
Isolated_Characters	Alef	222	134	44	44
	Beh	222	134	44	44
	Peh	222	134	44	44
	Theh	222	134	44	44
	She	222	134	44	44
	Jeem	222	134	44	44
	Cheh	222	134	44	44
	Heh	222	134	44	44
	Kheh	221	133	44	44
	Dal	222	134	44	44
	Zal	222	134	44	44
	Reh	222	134	44	44
	Zeh	222	134	44	44
	Jeh	222	134	44	44
	Seen	222	134	44	44
	Sheen	222	134	44	44
	Sad	222	134	44	44
	Dad	222	134	44	44
	Tweh	222	134	44	44
	Zweh	222	134	44	44
	Ain	222	134	44	44
	Ghain	222	134	44	44
	Feh	222	134	44	44
	Qaf	222	134	44	44
	Kaf	222	134	44	44
	Gaf	222	134	44	44
	Lam	222	134	44	44
	Meem	222	134	44	44
	Noon	222	134	44	44
	Waw	222	134	44	44
	Round-heh	222	134	44	44
	Heh-eyes	222	134	44	44
	Hamza	222	134	44	44
Yeh	220	132	44	44	
Te	222	134	44	44	
Heh- Hamza	222	134	44	44	
Waw-Hamza	222	134	44	44	
Isolated_Digits	Isolated_0	3965	2377	794	794
	Isolated_1	3277	1965	656	656
	Isolated_2	3288	1972	658	658
	Isolated_3	3214	1927	643	644
	Isolated_4	3023	1812	605	606
	Isolated_5	2908	1744	582	582
	Isolated_6	2910	1746	582	582
	Isolated_7	2858	1713	572	573
	Isolated_8	3074	1843	615	616

		Isolated 9	3668	2200	734	734
Numeral_Strings	Integer_strings	Lenght 2	2875	1725	573	577
		Lenght 3	1549	928	309	312
		Lenght 4	1300	778	260	262
		Lenght 6	1105	663	220	222
		Lenght 7	1100	658	221	221
	Real_Strings	Length 4	222	134	44	44
		Length 5	216	128	44	44
		Isolated Dot	411	245	83	83
Special_Characters	at the rate	215	127	44	44	
	Number sign	222	134	44	44	
	Afghani sign	221	133	44	44	
	Rupee sign	221	133	44	44	
	Riyal sign	222	134	44	44	
	Colon	221	133	44	44	
	Slash	221	133	44	44	
Words	One	222	134	44	44	
	Two	222	134	44	44	
	Three	222	134	44	44	
	Four	222	134	44	44	
	Five	222	134	44	44	
	Six	222	134	44	44	
	Seven	222	134	44	44	
	Eight	222	134	44	44	
	Nine	222	134	44	44	
	Ten	221	133	44	44	
	Twenty	222	134	44	44	
	Thirty	222	134	44	44	
	Forty	221	133	44	44	
	Fifty	222	134	44	44	
	Sixty	221	133	44	44	
	Seventy	221	133	44	44	
	Eighty	222	134	44	44	
	Ninety	222	134	44	44	
	Hundred	222	134	44	44	
	Thousand	222	134	44	44	
	100					
	Thousands	222	134	44	44	
	Million	221	133	44	44	
	Cash	222	134	44	44	
	Balance	222	134	44	44	
	Duty	220	132	44	44	
	Interest	222	134	44	44	
	Decrease	222	134	44	44	
	Received	222	134	44	44	
	Not					
	Received	221	133	44	44	
	Total	221	133	44	44	
	Amount	222	134	44	44	
Number	222	134	44	44		
Amount	222	134	44	44		
Item	222	134	44	44		
Delivery_1	222	134	44	44		
Period	222	134	44	44		

Article_1	222	134	44	44
Credit_1	222	134	44	44
Weight_1	221	133	44	44
Rent	222	134	44	44
Expire	222	134	44	44
Delivery_2	222	134	44	44
Weight_2	222	134	44	44
Width	222	134	44	44
Length	222	134	44	44
Cost	221	133	44	44
Price	222	134	44	44
Credit_2	222	134	44	44
Inches	222	134	44	44
Article_2	222	134	44	44
Sale	222	134	44	44
Debit	222	134	44	44
Product	222	134	44	44
Increase	222	134	44	44
Item	222	134	44	44
Payment	221	133	44	44
Dozen	222	134	44	44
Carton	221	133	44	44
Transfer	220	132	44	44
Plus	222	134	44	44
Gallon	220	132	44	44
Tons	222	134	44	44
Kilo	220	132	44	44
Milli	222	134	44	44
Centi	222	134	44	44
Gram	222	134	44	44
Liter	222	134	44	44
Meter	222	134	44	44
Inventory	222	134	44	44
Rupee	222	134	44	44
Afghani	222	134	44	44
Due	222	134	44	44
Stock	222	134	44	44

Table D.2: Detail Statistics of Pashto Database

Databases		Total	Training	Testing	Validation
Dates		216	128	44	44
Isolated_Characters	Alef	219	131	44	44
	Beh	219	131	44	44
	Peh	219	131	44	44
	The	219	131	44	44
	Tteh	219	131	44	44
	Theh	219	131	44	44
	Jeem	219	131	44	44
	Heh	219	131	44	44
	Kehh	219	131	44	44
	Zeem	219	131	44	44
	Tcheh	219	131	44	44
	Seem	219	131	44	44
	Dal	219	131	44	44
	Ddal	219	131	44	44
	Thal	219	131	44	44
	Reh	219	131	44	44
	Rreh	219	131	44	44
	Zeh	219	131	44	44
	Zzeh	219	131	44	44
	Geh	219	131	44	44
	Seen	219	131	44	44
	Sheen	219	131	44	44
	Kheen	219	131	44	44
	Sad	219	131	44	44
	Dad	219	131	44	44
	Twch	219	131	44	44
	Zway	219	131	44	44
	Ain	218	130	44	44
	Ghain	218	130	44	44
	Feh	218	130	44	44
	Qaf	218	130	44	44
	Kaf	218	130	44	44
	Gaf	218	130	44	44
	Lam	218	130	44	44
	Noon	218	130	44	44
	Noorr	218	130	44	44
	Waw	218	130	44	44
	Meem	218	130	44	44
	Yeh	219	131	44	44
	Strong-yeh	219	131	44	44
	Faminine Yeh	219	131	44	44
Soft-yeh	219	131	44	44	
Hamza-yeh	219	131	44	44	
Long-yeh	219	131	44	44	
Heh	218	130	44	44	
Hamza-hey	218	130	44	44	
Eyes-heh	217	129	44	44	
Hamza	219	131	44	44	

		Hamza-waw	217	129	44	44
Isolated_Digits		Isolated_0	3458	2074	692	692
		Isolated_1	3408	2044	682	682
		Isolated_2	3522	2112	705	705
		Isolated_3	3369	2019	675	675
		Isolated_4	3262	1958	652	652
		Isolated_5	3071	1843	614	614
		Isolated_6	3311	1985	663	663
		Isolated_7	3144	1886	629	629
		Isolated_8	3144	1886	629	629
		Isolated_9	3778	2266	756	756
Numeral_Strings	Integer_Strings	Lenght_2	2892	1752	570	570
		Lenght_3	1533	919	307	307
		Lenght_4	1313	787	263	263
		Lenght_6	1095	656	219	220
		Lenght_7	1095	656	219	220
	Real_Strings	Lenght_4	219	131	44	44
		Lenght_5	219	131	44	44
		Isolated_Dot	384	229	77	78
Specail_Symbols		at the rate	218	130	44	44
		Number sign	219	131	44	44
		Afghani sign	218	130	44	44
		Rupee sign	220	132	44	44
		Colon	219	131	44	44
		Slash	219	131	44	44
Words		Kilometer	218	130	44	44
		Meter	219	131	44	44
		Centimeter	216	128	44	44
		Millimeter	219	131	44	44
		Inches	219	131	44	44
		Gallon	218	130	44	44
		Liter	219	131	44	44
		Milliliter	219	131	44	44
		Tons	219	131	44	44
		Kilogram	219	131	44	44
		Gram	219	131	44	44
		Milligram	219	131	44	44
		Dozen	219	131	44	44
		Carton1	218	130	44	44
		Carton2	219	131	44	44
		One	219	131	44	44
		Two	219	131	44	44
		Three	218	130	44	44
		Four	219	131	44	44
		Five	219	131	44	44
		Six	219	131	44	44
		Seven	219	131	44	44
		Eight	219	131	44	44
		Nine	219	131	44	44
		Ten	219	131	44	44
		Twenty	219	131	44	44
		Thirty	219	131	44	44
		Forty	219	131	44	44

Fifty	219	131	44	44
Sixty	219	131	44	44
Seventy	219	131	44	44
Eighty	219	131	44	44
Ninety	219	131	44	44
Hundred	219	131	44	44
Thousand	219	131	44	44
100- Thousands	219	131	44	44
10 Millions	219	131	44	44
Decrease	219	131	44	44
Increase	217	129	44	44
Duty	219	131	44	44
Cash	219	131	44	44
Credit	219	131	44	44
Received	219	131	44	44
Not-received	219	131	44	44
Total	219	131	44	44
Number	219	131	44	44
Amount	219	131	44	44
Inventory	219	131	44	44
Item	219	131	44	44
Stock	219	131	44	44
Period	219	131	44	44
Article	219	131	44	44
Interest	219	131	44	44
Balance	219	131	44	44
Weight	219	131	44	44
Rent	219	131	44	44
Expire	219	131	44	44
Delivery	219	131	44	44
Tax	219	131	44	44
Volume	219	131	44	44
Width	219	131	44	44
Length	219	131	44	44
Due	219	131	44	44
Cost	218	130	44	44
Price1	219	131	44	44
Price2	219	131	44	44
Rupee	219	131	44	44
Afghani	219	131	44	44