Survey of Urdu OCR: An Offline Approach

Naila Fareen[1], Mohammad Abid Khan[2], Attash Durrani[1]
Dept. of computer science, Allama Iqbal Open University, Islamabad, Pakistan[1]; Dept. of Computer science, University of Peshawar, Peshawar, Pakistan[2]
nailafareen@ hotmail.com, mabid@upesh.edu.pk, attash.durrani@inksoft.net

Abstract

Optical Character Recognition (OCR) is the process of converting printed, handwritten and typed printed text into its equivalent machine readable form. Scanning and comparison techniques are considered to recognize printed text or numerical data. Once the scanned document is converted into machine readable form, the text can then be used in different applications, just like normal machine readable text. It saves time by not typing already printed material for data entry. OCR software attempts to identify characters by comparing figures to those stored in the software library. The discipline of OCR is an offspring of Pattern Recognition, Artificial Intelligence, and Computer Vision. Arabic script (having characters that are connected cursively) makes the recognition of Urdu text more difficult as compared to a language such as English having isolated characters when forming a word. In this research paper, an analysis of 8 years research papers (2002 to 2009) on Urdu OCR has been conducted to show the endeavors for the development of offline Urdu OCR covering both history and future work.

1. Introduction

There are many scripts available in Arabic and Urdu and many people have worked on Nastaleeq type. According to the best of the author's knowledge, none of the work is done on pattern recognition of Typed Urdu Naskh font in which the treasure of knowledge exist left by our predecessors. T. Nawaz and his/her co-authors discussed about the Urdu OCR Naskh font by using pattern matching techniques. The authors worked on pattern recognition for Unicode based computer fonts [1]. So a lot of attention is needed towards Urdu Naskh typed font.

The recognition of characters of Arabic script based languages is not an easy task because of its cursive nature. Arabic characters are connected even when printed or typewritten. The characters of Arabic script and similar other characters are used by a greater percentage of the world's population to write languages such as Arabic, Farsi (Persian), and Urdu. According to V. Margner and H. El Abed, research work on Arabic Optical Text Recognition increased considerably since the 1980's. First systems for Arabic printed text were available at the market in the 1990's. The above authors also did the collection of real world data. They performed scanning and labeling of the collected data to construct a database [2].

Arabic OCR for Printed characters was a research topic in 1990's and a comparison on published papers was reported in 2000 There are three Arabic OCR systems (Sakhr, IRIS, ABBY) that are available in the market but they do not suit for Urdu script, because Urdu has some additional characters. Therefore the OCR used for Arabic or Farsi will not accomplish all the needs for Urdu OCR [2].

2. Urdu OCR History

Urdu is the national language of Pakistan with around 180 million speakers. Urdu script belongs to the family of scripts based on Arabic script. It is a cursive script, i.e. individual characters are usually combined to form ligatures. Although many fonts are available for Urdu, the predominant fonts are Nastaleeq and Naskh. In order to automatically convert an Urdu document image into electronic form, an Urdu OCR system is needed. However, there has been very little work done in the area of Urdu OCR [3]. Urdu is a derivational word from Turkish and it means "horde" (Lashkar). Urdu is an Indo-European language of the Indo Aryan family [4].

Urdu computing started in early 1980s, creating multiple encodings each in different places, as a standard encoding scheme was missing at that time. With the advent of Unicode in early 1990s, some online publications have switched to Unicode, but much of the publications still continue to follow the traditional ad hoc encodings [5].

3. An Analysis of Offline Urdu OCR

On the basis of actual situation of research Table 1 gives an overview of recently published offline Urdu recognition systems and their accuracy.

Table 1: overview of offline Urdu text recognition system

Author(s)	Methodologies\Algorithms	Data Used	Results
S.A. Hussain & S. H. Amin (2002)	Multi-tier holistic approach and Feed Forward Back Propagation Neural Network	200 Carefully selected ligatures	100%
U.Pal & Sarkar (2003)	Water Reservoir Principle	3050 Characters	97.8%
F. Shafait, et al. (2006)	Geometric layout analysis system	Text line detection of 25 scanned images	Books 90%, Magazines 80% and news papers 72%
Z. Ahmed, et al. (2007)	Neural Networks	Old and newly written scripts used to evaluate results	93.4%
I.Shamsher, et al. (2007)	Feed Forward Neural Network	Ariel font Type in the Urdu alphabet set, 72pt font size	98.3%
A.Gulzar & S.Rehman (2007)	Omega: as typesetting Engine	7000 ligatures	Only 65 Nuqta clash
N.Shahzad, et al. (2009)	Subset of Rubine features, weighted and linear classifiers presented, to perform results	38*4 characters	Native Urdu participants, achieved: 92.8%, Non- Native Urdu participants achieved: 73%
T. Nawaz, et al. (2009)	Algorithm Used: Chain code calculation, segmentation, classification, character matching, testing, Unicode file creation.	Different printed Urdu text image files of different font size of Urdu isolated characters	89%
M.W. Sagheer, et al. (2009)	SVM, RBF and MN	3770 images of words.	98.61%
S.T. Javed et al. 2010	HMM and HTK	3655 Ligatures of Noori Nastaleeq	92%
S.S. Bukhari et al. 2011	Multiresolution Morphology and Ridge Based Method	25 Arabic and 20 Urdu Document Images.	99% for non text segmentation, 96% for Arabic and 92% for Urdu text-line Detection.

S. A. Hussain and S. H. Amin present a new approach of offline character recognition. For this purpose, they select the Noori Nastaleeq script by using Multi-tier Holistic approach for the recognition of a ligature [3]. According to U. Pal and A. Sarkar, writing style in Urdu is from right to left whereas it is from left to right in other Indian scripts [6]. Under the sub-heading “PREVIOUS WORK” the above authors mentioned, the accuracy achieved by their system is 97.8%, but it will not handle the messiness and variation of handwritten characters. It can be noted that an Urdu basic character may have four components. There is a structural similarity between Urdu and Arabic script. These authors presented their paper on the recognition of printed Urdu script. They use the water reservoir principles for the character recognition. They use the Hough transform technique for skew angle estimation [6].

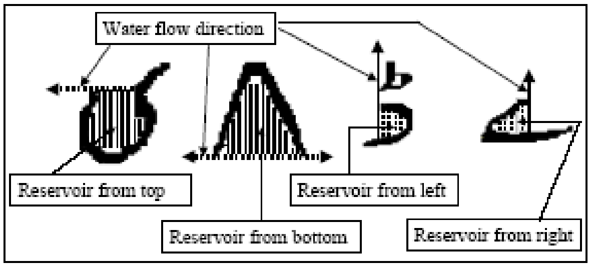


Fig.1. Different reservoirs and their water flow directions are shown in four characters. [6]

According to U. Pal and A. Sarkar, the proposed OCR system automatically detects individual text lines and then segments the character in each line. Their OCR system only recognizes the basic (isolated) characters. They actually want to explain about the complex and cursive nature of Urdu script. They tested the system on 3050 characters. They tested the prototype of this system on printed Urdu characters and currently achieved 97.8% character level accuracy on the average. They describe that Urdu alphabet consists of 39 basic characters [6].

According to F. Shafait and et al [7], layout analysis is the key component of an OCR system. They present a layout analysis system from Urdu document images for extracting text-lines in reading order. They are concerned with geometric layout analysis of Urdu documents. They present the block diagram which shows the flow of layout analysis as shown in the Fig 2.

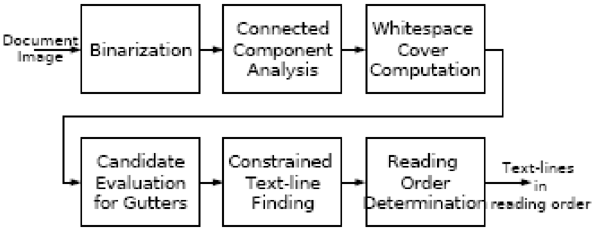

Document Image
Binarization
Connected Component Analysis
Whitespace Cover Computation
Candidate Evaluation for Gutters
Constrained Text-line Finding
Reading Order Determination
Text-lines in reading order

Fig 2: Block diagram of Layout analysis. [7]

According to these authors the results obtained on Urdu script are not as robust as those in the roman script. They collected 25 images of Urdu text to evaluate the performance of the described layout analysis system. They mentioned that due to the presence of dots and diacritics the results were not satisfactory as compared to Latin script. Mainly this paper analyzes, and to some extent detects, the layout of scanned documents. Their proposed algorithm achieved detection rate of more than 90% for line detection for the source taken from books and magazines, whereas in case of decreased inter-line spacing as in digest, decreased the detection results up to 80%. Newspapers source line detection was 72% [7].

4. Segmentation

Z. Ahmed and his/her co-authors describe that Urdu is written in different styles and shapes. They found that recognizing the connection of a word is not a big issue but understanding the shapes/forms of words become complex i.e. when a letter comes in the beginning, in the middle or at the end it changes its original shape. They recognized the printed Urdu script using neural networks and achieved 93.4% accuracy on the average. According to these authors, the main focus is on character segmentation. They mentioned that there is 58 character set defined by NLA (National Language Authority). They assumed that input script was diacritics free, where the diacritics in Urdu script distinguish the homonym easily [8].

M. I. Razzak and his co-authors try to combine the preprocessing steps for Online Character Recognition and present a novel technique for online preprocessing for removing the variation in both online and offline text. Whereas in printed offline approach, the variation rate is not high as found in handwritten documents. They mentioned that they performed different preprocessing steps on the input strokes, but they didn’t name those preprocessing steps. They proposed segmentation that was based on the threshold value and position with respect to the previous base character. They did smoothing on the chain code of the stroke. They transformed the input stroke into image to perform offline preprocessing steps. They claim “By using joint processing for online and offline OCR the efficiency can be increased”. But they didn’t mention the accuracy and recognition rate in percentage, so we are not fully aware about the achievements of this research [9].

M. Akram and S. Hussain presented the word segmentation for Urdu OCR system, they tested their model on the corpus of 150 sentences, and these sentences were composed of 2156 words and 6075 ligatures. They also mentioned that 65 unknown words and 2092known words. The identification rate of this model was 96.10% with 65.63% unknown words [10].

S. T. Javed et al. extracted the global transformational features from non segmented ligature. They used Hidden Markov Model (HMM) for recognition and HMM Tool Kit (HTK) to implement

HMM. For achieving 92% accuracy they tested the system with 3655 ligature of Noori Nastaleeq font [11].

S. S. Bukhari and his co-authors presents a robust layout analysis system from scanned Arabic script document images written in different languages (Arabic, Urdu, Persian) and styles (Naskh, Nastaliq) for extracting text-lines in reading order. They evaluated their system on 25 Arabic and 20 Urdu document images. They achieved 99% non text segmentation accuracy by using multi resolution morphology based method and above 96% text line detection accuracy for Arabic dataset and 92% for Urdu dataset by using ridge based method [12].

5. Recognition

I. Shamsher and his co-researchers presented that there is a lot of work done on the literature of Islamic studies and Urdu, which need to be transferred into electronic form. The above mentioned authors use their own proposed Feed Forward Neural Network Algorithm of MLP (Multi Layer Preceptrons) for the implementation of Urdu OCR Shown in the Fig 3. Their methodology consists of three layers, i.e. input layer, hidden layer and output layer. They worked on OCR system for printed Urdu. The accuracy rate at character level is 98.3% on the average. The software is tested in 72 pt font size only. This software, however, only recognizes the individual characters, so its scope is limited [13].

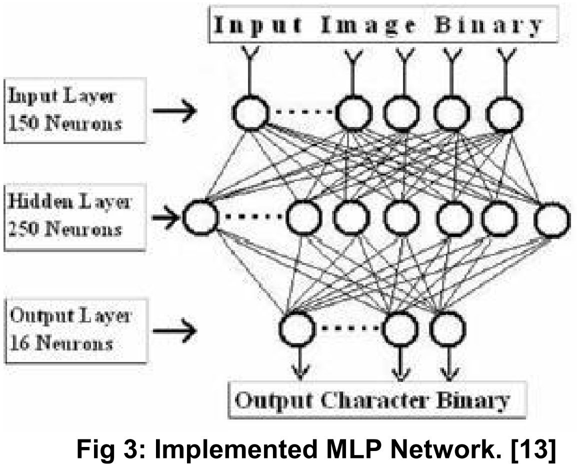
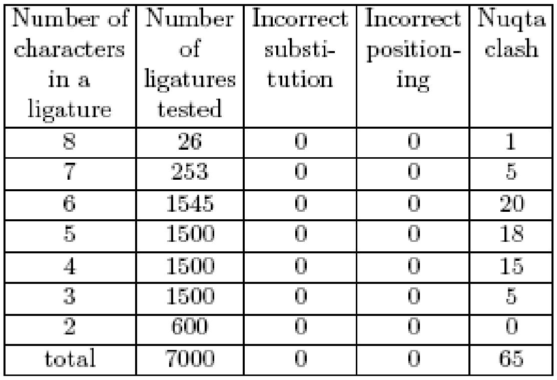


Fig 3: Implemented MLP Network. [13]

According to A. Gulzar & S. Rehman, Nastaleeq is a complex font. They discuss the complexity of Nastaleeq font and also provide the solution that uses Omega as the Typesetting Engine for rendering Nastaleeq. They discuss that there are more than 20,000 ligatures in Urdu and they use only 7000 ligatures randomly from the corpus of 20,000 valid ligatures [14]. Fig 4 shows the test results of A. Gulzar & S. Rehman.

Fig 4: Test results by using Omega. [14]



Number of characters in a ligature	Number of ligatures tested	Incorrect substi-tution	Incorrect position-ing	Nuqta clash
8	26	0	0	1
7	253	0	0	5
6	1545	0	0	20
5	1500	0	0	18
4	1500	0	0	15
3	1500	0	0	5
2	600	0	0	0
total	7000	0	0	65

N. Shahzad and his co-researchers in their paper "Urdu Qaeda: Recognition System for Isolated Urdu Characters" presented the online system for recognizing isolated, hand sketched Urdu characters. The system showed an accuracy of 92.8% for native Urdu writers. According to them, there is no significant research which has been directed towards the online recognition of the Urdu Language. They mentioned that “Urdu language consists of 38 basic characters”. The authors further say that most of the characters in Urdu language are multi-stroke. Each character has a single primary stroke and zero or more secondary strokes. They give the example of “ ش ”. In this case, according to these authors, “ س ” is the primary stroke and three dots are secondary strokes. However in the discussion section they have shown the “ ش ” which has single primary stroke and two secondary strokes as shown ش . So their earlier statement that “sheen (ش) is a multi-stroke character which consist of four strokes is not matched with the given image of sheen (ش). According to the above mentioned authors, some characters were not correctly recognized due to similarity in writing. They presented an example of similar characters which cause recognition problem as in the following case.

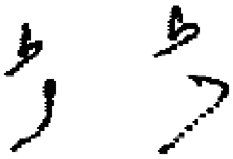
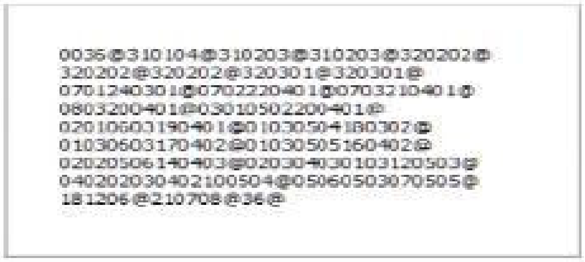


Fig 5: Similar Characters. [15]

These characters are similar in level of secondary stroke as the *toye-shosha* is similar, but the basic or primary characters are quite different. They only use the initial character of cursive script for recognition. They concluded that the system also recognizes a character that is not true in shape. The system should be able to reject the input as an invalid character so the user could learn and draw the character correctly [15].

J.Tariq and his co-workers develop a prototype of Urdu OCR which recognizes the isolated characters of Urdu language with the help of database, without using neural network and its accuracy rate is 97.43% [16].

T. Nawaz et al. discussed about the Urdu OCR Naskh font by using pattern matching techniques. These authors worked on pattern recognition for Unicode based computer fonts and proposed offline character recognition for isolated characters of Urdu language. They use the Chain code algorithm for character matching [1]. Calculating alternating on and off pixels the chain code is generated as shown in the Fig 6.



```
0036@310104@310203@310203@320202@
320202@320202@320301@320301@
0701240301@0702220401@0703210401@
0803200401@03010502200401@
0201060319040 1@01030504180302@
01030603170402@01030505160402@
02020506140403@020304030103120503@
04020203040210050 4@05060503070505@
181206@210708@36@
```

Fig 6: Calculated String for chain code. [1]

There are many words in Urdu, which are formed using only isolated characters. For example, see Fig 7.

وہ , اور , ا ر د و

Fig 7: Urdu Isolated Characters

But these are only few words, so the claim made in T. Nawaz's paper that "Urdu language forms words by combining isolated characters" is partially correct, as examples mentioned above. They also mentioned, "Urdu is a cursive language, having connected characters making words". The examples of such words are in Fig 8.

پاکستان , اسلام آباد , تحقیق , نسخ

Fig 8: Urdu Ligatures

They worked only for isolated characters, so the scope for their research is limited. The majority of Urdu words are connected characters. They claimed about accuracy up to 89% with the rate of 15 chars/sec. They also mentioned, "Urdu character set has 40 characters". But there is a controversy over the number of letters in Urdu alphabet. National Language Authority declared 58 letters of Urdu, in a meeting dated 26 January 2004 [17].

As per Dr. Rauf Parekh "The controversy over the total and correct numbers of letters in Urdu alphabet has been running for over 200 years now". The algorithm chain code calculation which is normally used in such OCR systems to recognize characters with the segmentation i.e. line segmentation, character segmentation and two levels segmentation [18] .

A. Durrani describes the shapes of four letters as given in Fig 1 that have become the reason of segmentation or space between words. Wherever any of these four letters appear, the ligature of work will break. There is a rule in Urdu that all letters become combined until the word is finished and the last letter will be in full shape [19].

ا ر د و

Fig 9: Letters of the word Urdu

Urdu characters in isolated shape have only one form but in case of connected, their shapes change ranging from 2 to 4 as given in Fig 10.

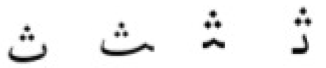


Fig 10: Urdu character: Shapes

There are many fonts available in Arabic and Urdu but it is considered that "Naskh" and "Nastaleeq" are the main two fonts, Nastaleeq has some complexities. Many people have worked on Nastaleeq type. Calligraphy of Nastaleeq is still waiting for research projects. In Pakistan, none of the work is done on pattern recognition of Typed Urdu Naskh font.

M.W. Sagheer and his co-authors worked on large database for off-line handwriting recognition. They conducted the recognition of Urdu digits and achieved the accuracy of 98.61%. For the recognition process they use the methodologies of Support Vector Machine (SVM), Radial Base Function (RBF) and Moment Normalization (MN) [20].

6. Conclusion and Future Work

This paper describes research on Urdu OCR. It has discussed methods of OCR and classified them according to different criteria. It is the first Urdu offline character recognition survey to give testing procedures and recognition rates for as many systems as possible. However, current systems are applied to restricted domains and/or have only been tested on small datasets.

Future research and testing are needed to develop systems for widespread use. Considering all the aspects in the previous section, the next step is to provide better offline Urdu OCR for the typography and pattern recognition. So a lot of attention is needed towards Urdu Naskh to bring it in a working stage, especially for the type cast by type foundries the pages partitioned accordingly. Urdu typography and calligraphy are enormously different fields but most of the people mixed-up both fields, so the distinction between them may be considered. The work on Urdu typography of Typed Urdu Naskh font that was developed and used by foundries before the advent of computerized printing has not been touched by any researcher. So there is dire need to convert that work into electronic form so that everyone can get benefit from the work that has almost been diminished.

References

Articles in journals:

[1] T. Nawaz, S. A. Naqvi, H. Rehman and A. Faiz "Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique", International Journal of Image Processing, 3(3), pp. 99-104, 2009.

Articles in conference proceedings:

[2] V. Margner and H. El Abed, "Arabic Word and Text Recognition -- Current Developments ", *2^{nd} International conference on Arabic Language Resources and tools,* Cairo, Egypt, pp.31-36, April 2009

[3] S. A. Hussain, and S. H. Amin, "A Multi-tier Holistic approach for Urdu Nastaliq Recognition", *IEEE INMIC,* Pakistan. 2002.

[4] W.Anwar , X. Wang and X.L. Wang. "A Survey of Automatic Urdu Language Processing" Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 13-16, August 2006.

[5] S. Hussain, "Resources for Urdu Language Processing", *the 6th Workshop on Asian Languae Resources*, 2008.

[6] U. Pal and A. Sarkar. "Recognition of Printed Urdu Script", *ICDAR , IEEE*, 2003.

[7] F. Shafait, Adnan-ul-Hassan, D. Keysers and T. M. Breuel "Layout Analysis of Urdu Document Image", *Multitopic Conference,, INMIC'06. IEEE*, 05-07. 2006.

[8] Z. Ahmed, J. K. Orakzai, I. Shamsher and A. Adnan "Urdu Nastaleeq Optical Character Recogniotion", *World Academy of Science, Engineering and Technology,* (32), pp.249-252, 2007.

[9] M. I. Razzak, S. A. Hussain, M. Sher and Z. S. Khan "Combining Offline and Online Preprocessing for Online Urdu Character Recognition", *Proceedings of the International MultiConference of Engineering and Computer Science*, (1). ISBN: 978-988-17012-2-0., 2009.

[10] M. Akram and S.Hussain, "Word Segmentation For Urdu OCR System",Proceeding of the 8th Workshop on Asian Language Resources,Beijing, China, August,2010.

[11] S. T. Javed, S. Hussain, A. Maqbool, S.Asloob, S. Jamil and H. Moin "Sagmetation Free Nastalique Urdu OCR " *Word Academy of Science, Engineering and Technology,* 70, 2010

[12] S. S. Bukhari, F. Shafait and T. M. Breuel "High Performance Layout Analysis of Arabic and Urdu Document Images" 11th *International Conference on Document Analysis and Recognition, ICDAR'11*. Beijing, China, September 2011.

[13] I. Shamsher, Z. Ahmad, J. K. Orakzai, and A. Adnan "OCR For Printed Urdu Script Using Feed Forward Neural Network", World Academy of Science, Engineering and Technology,34,2007.

[14] A. Gulzar and S. Rehman. "Nastaleeq : a Challenge Accepted by Omega", *TUGboat, XVII European TEX Conference*, 29(1), pp.89-94, 2007.

[15] N. Shahzad, B. Paulson and T. Hammond, "Urdu Qaeda: Recognition System for Isolated Urdu Characters" IUI 2009 Workshop on Sketch Recognition, Sanibel Island, Florida, February 8, 2009

[16] J. Tariq, U.Nauman and M.U.Naru "A novel approach to construct OCR for printed Urdu isolated characters", Second international conference on Computer Engineering and Technology (ICCET), 3, pp. 495-498, June 2010.

Electronic Resources:

[17] National Language Authority, "Urdu Alphabet", [On line: *www.nla.gov.pk],* Retrieved: (07.05.2010).

[18] Rauf Parekh, "Controversy over number of letters in Urdu alphabet", Dawn (English Newwspaper), Karachi/Islamabad, Pakistan, (07.15.2009).

Books:

[19] A. Durrani. "Urdu Informatics", *National Language Authority,* Islamabad, Pakistan*,* 2008

[20] M. W. Sagheer, C. L. He , N. Nobile and C. Y. Suen " A New Large Urdu Database for Off-Line Handwriting Recognition" *Image Analysis and Processing – ICIAP 2009,* Springer Berlin / Heidelberg , pp. 538-546, 2009.