Optical Character Recognition Techniques in Urdu- A Survey

Vippon Preet Kour
Department of Computer Science & Engineering
SMVDU, Katra

Naveen Kumar Gondhi
Department of Computer Science & Engineering
SMVDU, Katra

ABSTRACT
The survey of the optical character reader for Urdu like cursive languages is based on the various techniques and studies performed on designing and implementation of the optical character reader. As, Urdu language has Nastaliq font so different approaches were applied on this font so as to get the desired result. Survey is being performed on all the techniques whether segmentation based, on-line or off-line etc, then all the data gathered is represented in a tabular manner so as to make it an ease to understand or to have an idea of the concept by visualizing the table at once. Non existence of the Urdu OCR has limited the concept a digital Urdu library and this nonexistence leads a pathway for immense research in this field.

General Terms
Pattern Recognition, Nastaliq font, Urdu language, Pushto language.

Keywords
Image Segmentation, Optical Character Reader, Feature Extraction, Classification.

1. INTRODUCTION
The character recognition or the optical character recognition is the process of the mechanical or electronic conversion of the type images of handwritten text or printed text into machine encoded text whether the text is taken from a handwritten document, photo of the document. It is a method of digitizing the printed texts so that they can be edited, searched, stored, more compactly and can be displayed online using machine processes. The applications of the OCR's are mostly for the blind impaired users, data entry for business documents, number plate recognition, defeating CAPTCHA in anti-bot systems. The OCR works with implementation of various techniques, tools and on the basis of them the accuracy level is obtained and thus a result is generated.

There are a variety of languages spoken all over the world. Many people are multi-linguistic, but certain languages are having immense cultural influence from the prehistoric times. The main languages of the era are Hindi, Sanskrit, Urdu, Punjabi, Sindhi, Pushto etc. Urdu is the national language of Pakistan, and an official language of six states of India. It is also one of the official languages recognized in the Constitution of India. Urdu language has some specialized vocabulary and apart from it Urdu is intelligible with standard Hindi. In the Asian continent, the Urdu language came under the influence of **B**ritish rule, when they replaced the Persian language by the Urdu language. Urdu language writing style basically comes under the 'Cursive Languages Writing Style'. Urdu is written in Nastaliq format. As there is a lot of old, popular literature written in Urdu language present in handwritten form, but all this is not digitized. So, due to the desire, demand and popularity of the Urdu literature, there is a need to develop and design an Optical Character Reader for Urdu Language. The objective of character recognition is to imitate the human reading ability, with the human accuracy but at far higher accuracy.

2. CHALLENGES IN URDU SCRIPT
There are various challenges in Urdu like cursive scripts and they are

2.1 Bi-directionality
Urdu language is written bidirectional, as the characters are written are written from right to left while the numerals are written left to right. So, this makes an OCR design much chaotic and complex.

2.2 Non-Monotonicity
The writing pattern of Urdu is quite different from the other cursive languages. In the Urdu like scripts, one frequently goes back to the already written character as certain letters consist of stoke that goes back and beyond the previous character e.g.; JEEM, HAAY.

2.3 Context sensitivity
Each character in Urdu language changes its shape in accordance with the neighboring character. Thus a character can have different shapes.

2.4 Complex dot placement
There is always some dot displacement in Urdu language depending upon the character presence with the neighboring characters. Due to this the dots displace from their standard positions**.**

2.5 Spacing
Like the English language, there is no definite procedure to understand the intra word spacing in case of Urdu language. This makes the readability a bit difficult task.

2.6 Looping
In Urdu language certain characters have negative and positive loops. The negative loops form shape resembling to the circle or oval or vice versa**.**

میں زمانہ اگلے ہیں کہتے

In this example there are both positive and negative loops. For example: meem and gaaf.

3. URDU WRITING STYLES
The various Urdu writing styles are present based on the past styles followed by different writers and people around the globe and they are

3.1 Urdu Script

It is an extension of the Persian alphabet from right to left, which itself is an extension of the Arabic alphabet. This feature is known as the Persian Calligraphy, but the Urdu language follows the Nastaliq style of this calligraphy. It is the most popular and commonly used style of the Urdu language used so far.

3.2 Kaithi Script

This script was used in the British administration courts of Bengal, Bihar, North-West provinces and Oudh. It is a highly Persianized and technical form of the Urdu language and was used prominently in the 19th century. Most of the legal operations or paperwork of the British India was executed in this script.

3.3 Devanagari Script

The introduction of orthographic features by publishers into the Devanagari script solved the purpose of representing the Perso-Arabic etymology of Urdu words. This is the most popular script adopted for publishing journals and other technical tasks.

3.4 Roman script

Due to the ease and availability of Roman movable type of printing press, Urdu was occasionally written in Roman script. It is prominently used over internet as well by the youngsters

Fig.1: The Urdu Nastaliq alphabets with their names Devanagari and Latin alphabets.

4. OPTICAL CHARACTER RECOGNITION TYPES

Based on the type of the input mode, the OCR can be classified as On-line OCR and Off-line OCR. The online character recognition system consists of five components.

4.1 Image acquisition

This step involves binarization, filtering, smoothing, slant correction, skew detection, and thinning and baseline detection to improve the performance. This step affects the reliability and efficiency.

4.2 Preprocessing

This step involves the tasks such as the separation of dots touching the base of the ligature.

4.3 Segmentation

In this step, the text of a paragraph is segmented into lines and then the lines into words and the words into sub words/ligatures. This step has two kinds of approaches viz "segmentation free or holistic methods" and "segmentation based or analytical methods".

4.5 Feature Extraction

This step involves the extraction of unique and salient patterns from the input image to enhance the discrimination power and reduce data for the classification. The extracted features can be classified as Statistical features, Global features and. Structural features.

4.6 Recognition/classification

On the basis of the extracted features the classification /recognition is the main decision making stage. The pattern is identified and recognized from the input features.

5. STEPS OF AN OCR PROCESS

The diagramatic flow diagram shows how and what main steps are taken into consideration while designing an OCR, and they are as shown in the fig:

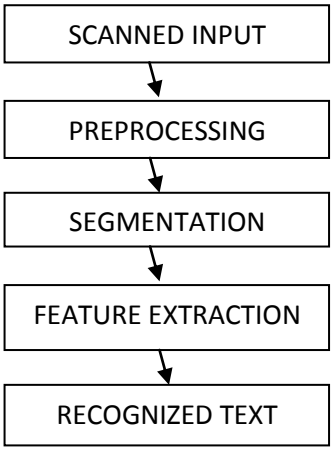

Fig. 2: Steps of OCR for cursive languages:

A brief description of the diagram is given below as:

5.1 Input or scanned pages

This is the data which needs to be recognized by the optical character reader. This can be in any form i.e., either in text form or in the form of an image.

5.2 Pre –processing

This step is performed on raw data to prepare it for another processing procedure. It is the preliminary step that transforms the data into a format that can be more effectively processed with ease. It is an important step and usually consists of binarization, filtering, smoothing, slant correction, skew detection, thinning, baseline detection etc. This requires fineness while carrying out the tasks as it can severely and adversely affect the upcoming steps.

5.3 Segmentation

It is defined as the process by which a given data is divided into sub data or we can say that in order to detect what actually is contained in the image or input, the division of the

given image or data is done into subparts and then taking these subparts after processing are merged together to determine what actually was in the input, or the subparts can be feed to the next step for processing.

5.4 Feature extraction

This step starts from the initial step of measured data and extracts the derived or desired features or values which are intended to be non-redundant and informative. The extracted or selected features are assumed to contain the desired information or data. It extracted features can be classified as Structural features, Statistical features, Global transformations.

5.5 Recognized text

This is the final output result or desired data that has been obtained from the application of the previous steps.

6. APPROACHES

For the design of an optical character reader for the Urdu language, a wide variety of techniques/approaches were used by different researchers.

6.1 Segmentation based approaches

Segmentation is the process in which the given data is divided i.e., the subparts of a particular data are formed. So the data that is to be segmented can be in any form e.g., a paragraph, a line, a word etc. The subpart or the divisible part is made such that it can be processed easily in the upcoming steps. The segmentation is followed in such a sequential manner that the paragraph is segmented into the individual lines and the line is then segmented into different words and at last the words are segmented into the characters or the alphabets.

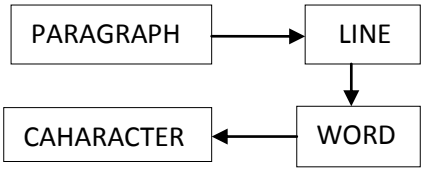

Fig.3: Step wise representation of segmentation

The commonly two types of approaches in segmentation include Segmentation free or holistic approach and Segmentation based or analytical approach. The analytical approach is further of two methods i.e., indirect and direct. In direct method, the ligatures are not further segmented while in indirect method, a word is separated directly into the letters using a number of heuristics that identify all of the segmentation points of a character. A ligature is resolved by splitting it into smaller elements that might be letters or less than letters such as sub letters or small stokes which further need segmentation. Hence, segmentation proves to be an extensive and important approach in OCR.

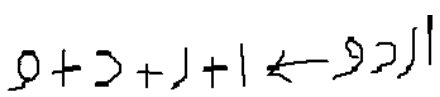

Fig.4: Segmented Urdu word

6.2 Hidden Markov Models (HMM)

It is a statistical model in which the system being modeled is assumed to be a process with hidden states. The dynamic Bayesian Network is used for representing HMM in the simpler form. It is based a little on the forward-backward procedure of a optimal non-linear filtering problem. In this model the state is not directly visible to the observer but the output dependent on the state is visible. Each state has the probability distribution over the possible output tokens and the sequence of tokens generated by the model give information about the sequence of states. The 'hidden' here indicates a state through which the model passes but not the parameters of the model. The HMM are generally applied in temporal pattern recognition such as speech, handwriting, gesture recognition, musical score following etc. The diagrammatic representation of the model is shown as:

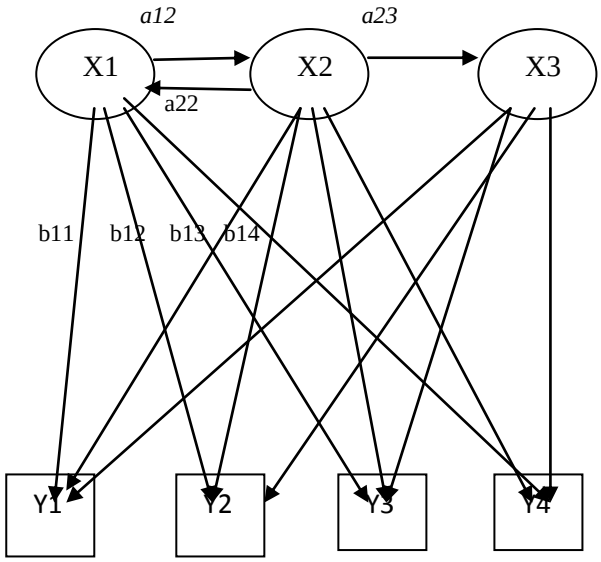

Fig.5: Hidden Markov Model with X-States, Y-Possible observations, a-State transition probabilities, b-Output probabilities.

6.3 Template matching:

In this approach we find the small parts of an image which match the template image. It is a feature based approach i.e., strong features are taken into consideration while matching. If the strong features are not present, then the template based approach is followed and it proves to be effective. It requires sampling of a large number of points and then from that sample we start matching the points. If the template may not provide a direct match, then the other methods like motion tracking and occlusion handling are performed. This approach has various applications as face detection, visual object recognition, car plate recognition etc. It is a simple method used for classification and pattern recognition. A database of templates is used for matching which is called the training data. It can identify scanned or computer written characters, numbers and the secondary characters are known as diacritics. The template image is moved to all possible positions of the source image and an exact match with nearest representation is extracted and taken into consideration. In most of the cases all matching is done on pixel by pixel basis.

6.4 Unicode Mapping

It is a computing industry standard for the consistent encoding, representation and handling of text in most writing systems. The latest version of Unicode contains a repository of more than 120,000 characters covering about 129 modern and historic scripts along with multiple symbol sets. Unicode can be implemented by different character encodings. All of the possible sequences of segments are generated and stored in a file. After the segments are generated, we produce Unicode of each segment. One character can have multiple segments (over segmentation), while others can combine to

construct one segment (under segmentation). One character can have different number/types of diacritics to distinguish among characters. So, a state machine has been developed to cope up with all the combinations. One sequence of states exhausts inputs and returns the Unicode value of one character. When a ligature is tested, long sequence for all segments is generated and then found in the state table. If any fulfilling sequence is found then the value is acquired, otherwise one state from the sequence is dropped and again searched in the table. So, it is the longest sequence matching algorithm.

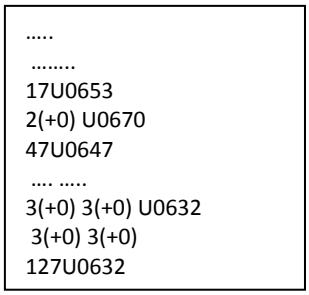

Fig.6: State Transition Table

Table 1: Comparison of various OCR approaches

Approach	Authors	Data Sets	Classification	Accuracy
Segmentation	Hussain et al[1]	200 Ligatures	Connected component labelling and centroid to centroid distance	100%
	Pal et al[2]	Small variety Characters	Horizontal and vertical profile, component labeling	96.90%
	Akram and Hussain[3]	150 sentences composed of 6075 ligatures and 2156 words	Ligature used as a structural method, trigram trained on co-occurrence information of ligatures and words in the corpus.	99.40%
Isolated Character Recognition	Pal et al[2]	3050 characters	Topological, contour and water reservoir, template matching.	98%
	Zaman et al[4]	106 tures	Pixel values using row major and column major order, template matching	95.00%
Machine Printed Ligature Or Word Recognition	Hussain et al[1]	200 Ligatures	FFNN Back, solidity, number of holes, axis eccentricity, moments.	100%
Handwritten Isolated Character, Ligature, Word And Numeral Recognition	Sagheer et al[5]	60329	Numeral, gradient, SVM	99%
	Basu et al[6]	3000	Numeral, QTLR	96.20%
Online Isolated Character, Ligature Or Word And Numeral Recognition	Sardar[7]	1050	Sliding window and HMM, KNN	97%
	Razzak et al[8]	900 Images	Numeral, Structural ,Rule Based	96.30%
	Razzak et al[9]	900 ligatures	Numeral, Fuzzy Logic, Fuzzy rule, Hybrid and HMM	97.80%

7. CONCLUSION

The reliable Urdu script OCR is still a far cry due to immense challenges. In particular the Nastaliq style of writing and its geometrical difference from the Naksh style of writing makes this more challenging. The researchers had tried both online and offline forms of the handwritten text, but haven't yet been more successful in either of them. Isolated and ligature based recognition for the Urdu script is more enthusiastic parameter in research so far. Till this date there is no multilingual OCR available, but there is a need to develop algorithms that can incorporate unlimited database as there is high similarity among Arabic script languages.

8. REFERENCES

[1] S.A. Husain, A multi-tier holistic approach for urdu Nastaliq recognition, in: Proceedings of the 6th International Multitopic IEEE Conference (INMIC'02), 2002.

[2] U. Pal, A. Sarkar, Recognition of printed Urdu script, in: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003.

[3] M. Akram, S. Hussain, Word segmentation for urdu OCR system, in: Proceedings of the 8th Workshop on Asian Language Resources. Asian Federation for Natural Language Processing, Beijing, China, 2010.

[4] S. Zaman, W. Slany, F. Sahito, Recognition of segmented Arabic/Urdu characters using pixel values as their features, in: Proceedings of the 1st International

Conference on Computer and Information Technology (ICCIT'2012), 2012

[5] M.W. Sagheer, C.L. He, N. Nobile, C.Y. Suen, A new large Urdu database for off-Line handwriting recognition 5716 (2009).

[6] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, D.K. Basu, A novel framework for automatic sorting of postal documents with multi-script address blocks, Pattern Recognition 43 (10) (2010) .

[7] S. Sardar, A. Wahab, Optical character recognition system for Urdu: online and offline OCR irrespective of fonts, in: Proceedings of the International Conference on Information and Emerging Technologies (ICIET), Karachi, Pakistan, 2010.

[8] M.I. Razzak, A. Belaïd, S.A. Hussain, Effect of ghost character theory on arabic script based languages character recognition, in: Proceedings of the WASE Global Conference on Image Processing and Analysis (GCIA'09), Taiwan, China, 2009.

[9] M.I. Razzak, F. Anwar, S.A. Husain, A. Belaïd, M. Sher, HMM and fuzzy logic: a hybrid approach for online urdu script-based languages' character recognition, Knowledge Based Systems 23 (8) (2010)

[10] S.T. Javed, Investigation into a segmentation based OCR for the Nastaleeq writing system (Master's thesis). National University of Computer & Emerging Sciences, Lahore, Pakistan, 2007.

[11] Z.A. Shah, Ligature based optical character recognition of Urdu-Nastaleeq font, in: Proceedings of the 6th International Multitopic IEEE Conference (INMIC'02), 2002.

[12] S.T. Javed, S. Hussain, Improving Nastalique-specific pre-recognition process for Urdu OCR, in: Proceedings of the 13th International Multitopic IEEE Conference (INMIC'09), 2009.

[13] S.F. Rashid, S.S. Bukhari, F. Shafait, T.M. Breuel, A discriminative learning approach for orientation detection of urdu document images, in: Proceedings of the 13th International Multitopic IEEE Conference (INMIC'09), 2009

[14] M. Riley, Beyond quasi-stationarity: designing time-frequency representation for speech signals in : Proceedings of the International Conference on Acoustics Speech and Signal Processing(ICASSP87), vol. 12, 1987 ,pp, 657-660.

[15] Nabeel Shahzad,Brandon Paulson and Tracy Hammond Urdu Qaeda: Recognition System for Isolated UrduCharacters IUI 2009 Workshop on Sketch Recognition February 8, 2009, Sanibel Island, Florida Chair: Tracy Hammond

[16] Tabassam Nawaz, Syed Ammar Hassan Shah Naqvi, Habib ur Rehman & Anoshia Faiz Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique International Journal of Image Processing, (IJIP)Volume (3) : Issue (3)

[17] Sohail Abdul Sattar Shams-ul Haque Mahmood Khan Pathan "A Finite State Model for Urdu Nastalique Optical Character Recognition ",IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.9, September 2009

[18] Faisal Shafait, Adnan-ul-Hasan, Daniel Keysers, and Thomas M. Breuel, "Layout Analysis of Urdu Document Images," [Multitopic Conference, 2006. INMIC '06. IEEE,p. 293 – 298.]

[19] S.A.Hussain, Anwar F., Asma. "Online Urdu Character Recognition System." MVA2007 IAPR Conference on Machine Vision Applications.

[20] Liana M & Venu G. (2006). Offline Arabic Handwriting Recognition: A Survey. IEEE,Transactions On Pattern Analysis and Machine Intelligence, vol. 28, No. 5, pp. 712-724.I.

[21] R. Safabakhsh and P. Adibi. (2005). Nastaaligh Handwritten Word Recognition Using a ContinuousDensity variable-Duration HMM. The Arabian J. Science and Eng., vol.30, pp. 95-118.