Optical Character Recognition System for Urdu.
Online and Offline OCR Irrespective of Fonts.

Abdul Wahab[1] and Syed Najam ul Haque[2]
SZABIST
Karachi, Pakistan

Abstract: *The subject of Urdu OCR is of real importance. There are about 60 to 80 million speakers of Urdu language ranked as fifth most spoken language with 4.7 percent of the total world population, spoken in South Asia vastly in Pakistan and India . Huge amount of valuable Urdu literature from philosophy to sciences is in vanishing and useless form because it is has not been digitized till now. More importantly many of the native speakers of Urdu, especially in Pakistan can only read and write Urdu language and very rare data is available for them on internet and in digitized form. Because of its complexity very rare and partially research work and implementation has been done and therefore no complete OCR for Urdu language exits till now. More importantly mostly research has been done for Urdu OCR is with respect to scripts, fonts and text environment which are other obstacles in the way of making complete OCR. So we have research and moderately implements online and offline OCR system which is irrespective of Urdu scripts and fonts.*

Keywords: *OCR, Segmentation, Pre-processing, Line Extraction, Primary Stroke, Secondary Stroke, Component Extraction.*

INTRODUCTION

Many languages of the world are being adopted for different languages. Base language character sets serves as scripts in adopted language. Urdu language is also one of the languages which contain the features, properties, scripts and writing styles of duo languages Arabic and Persian. Arabic is the base language from which the Persian is take up and Urdu is espouse from both Arabic and Persian. Its base is also from Arabic but also contains features form Persian.

The most prevalent Urdu script is blend of Naskh, Arabic style and Talique, Persian style called Nastalique script. [1]

Urdu language contains both properties of Naskh and Talique. It is cursive in nature, which makes it more difficult for conventional algorithms to work on it. [2]

Urdu language has 39 characters which exist in 2 – 4 subsets, depending on the occurrence of the character; isolated, initial, middle and final.

Character recognition for any script lies in between the handwriting and printed character of that script. Worst character features can be obtained from handwriting and most high-quality features can be extracted from printed material.

Online character recognition is a process which is used for handwriting recognition, example digital pens. Online character recognition required real time data from user. Offline character recognition can have both handwriting and printed material. Offline character recognition mostly used for printed papers, book etc.

Preprocessing is the most vital stage in the Urdu OCR; it acts as a backbone for consistency and efficiency of next stages feature extraction and recognition. Preprocessing consists of text area extraction, text line extraction, baseline detection, component segmentation, character segmentation, primary and secondary stroke extraction. Urdu words consist of two or more characters which are connected through an imaginary line called baseline. Most of the researchers have use the methodology in which character segmentation is neglected and whole ligature is send for recognition.

Feature extraction occupy central role for the accuracy of recognition. In this stage structured information which is more related to writing like dot, loops and branches are computed.

Statistical information is also gathered which include numerical measurements and computing over image. [3]

Recognition is the last stage in OCR in which character or ligatures are recognized by there features. Many schemes have been adopted by researchers majority have implemented Artificial Neural Network for this purpose.

The fundamentals of this research is to propose methodology that works for both online and offline character recognition. Hence we introduce the algorithms that work for both handwriting and printed material irrespective of fonts and scripts, so the obstacles in making the complete Urdu OCR can be diminish.

PREVIOUS WORK

Optical character recognition has always been sparkling area for researchers over a long time for different languages. However for Urdu OCR not much research has been done [4].

Previous researches have been done in particular domains each one of them has their different point of views. Previous research on OCR can be classified into three main categories: preprocessing, features extraction and recognition [5].

Preprocessing stage includes normalization of words where image is often converted in to more brief representation prior to recognition.

Faisal Shafait et.al [6] evaluates an existing system on document analysis for Roman script text on Urdu documents and describes its methods and the main changes necessary to adapt it to Urdu script.

Pal et al [7] proposed a system for printed Urdu script, in which they perform skew correction, line segmentation by horizontal projection, component extraction by vertical projection. Finally, the ligatures are put forward to the OCR system for recognition. S.A.Husain et.al [8] performs 2 to 3 pixel de-hooking and smoothing in preprocessing phase.

S. Mozaffari et al [9] uses skeleton method in which character is represented as one-pixel thick which shows only centerlines of the character. Another method is used by R. Safabakhsh & P. Adibi [10] in which absolute position of the first pixel and the relative positions of successive pixels along the character's border are stored.

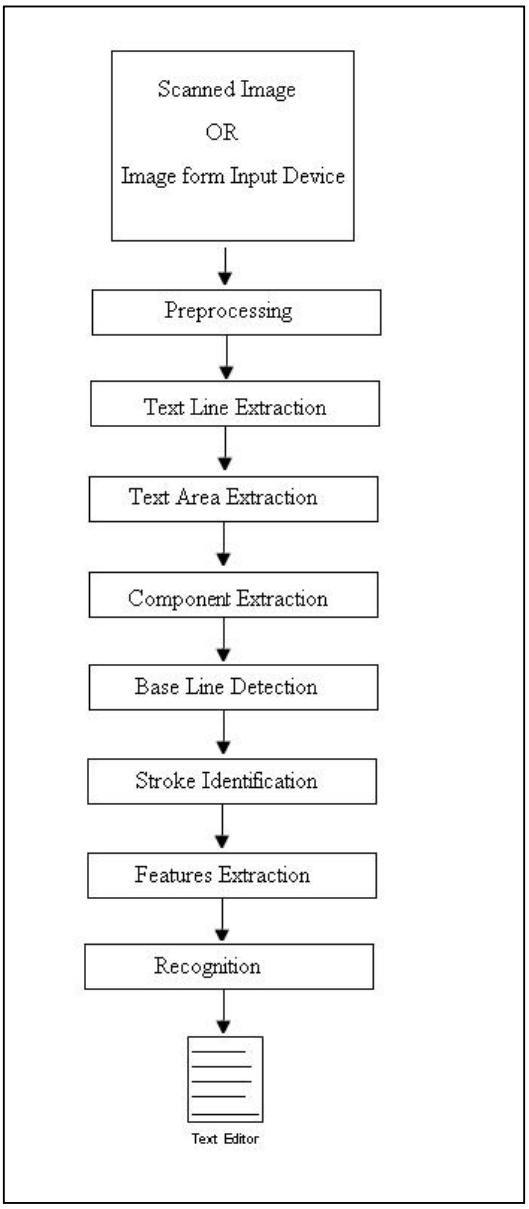
Features extraction comprise of structured and statistical computing. It involves the feature particularly related to writing styles and statistical information of character or ligatures.

S. A. Husain et.al [11] used vector containing 20 syntactical features. They identified loop, intersection, writing styles of ligatures. Syed. Afaq Husain et.al [12] used Solidity, Number of Holes, Axis Ratio, Eccentricity, Moments, Normalized segment length, curvature, ratio of bounding box width and height in feature extraction.

Tabassam Nawa et.al [13] used chain code for character recognition. Nabeel Shahzad et.al [14] used length and angel of the bounding box diagonal, distance, sine and cosine of the angle between first and last point, length of the primary stroke, angle traversed, absolute value of angle at each point as features of characters.

Character recognition is the final stage in which recognition algorithms are applied to segmented images with along with the features extracted.

Sohail Abdul Sattar et.al [15] presented a finite state model for character recognition. Inam Shamsher el.al [16] used Feed Forward neural network for character recognition.



Syed. Afaq Husain et.al [17] used Feed Forward Back propagation neural network for character recognition. Nabeel Shahzad et.al [18] used the weighted linear classifier for recognition.

PROPOSED SYSTEM

OCR system for Urdu language is mean to be complex because of the complication in Urdu writing styles and scripts. We have proposed methodology that works on both online and offline OCR irrespective of fonts or scripts. We have used segmentation free approach in which only ligatures are segmented.

Block Diagram of Proposed System

Preprocessing

Preprocessing involves binaraization, smoothing and noise analysis. We have applied gray scale threshold for converting images into binary form.

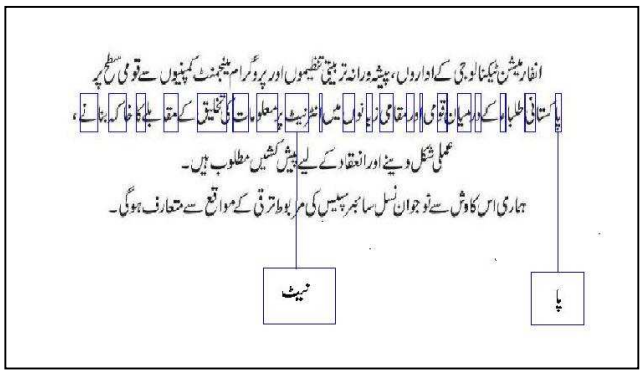
Text Line Extraction

As Urdu is written from left to right we have examined the binary image form left top to right bottom. Algorithm read pixels from left to right and stored the vertical positions where black pixels are not found which are just after the text line.

18 فروری الیکشن کے روز تک عدالتوں کے مکمل بائیکاٹ کا اعلان کیا۔ ملک بھر میں وکلاء نے مقامی سطح پر بھی
احتجاجی مظاہرے کیے اور عدالتوں کا بائیکاٹ جاری رکھا۔ تفصیلات کے مطابق اسلام آباد میں وکلاء، سول سوسائٹی
کے نمائندوں اور تحریک انصاف کے کارکنان اظہار یکجہتی کیلئے سپریم کورٹ بار کے نظر بند صدر اعتزاز احسن کی
رہائش گاہ گئے جہاں انہوں نے حکومت کے خلاف شدید نعرے بازی کی۔ بعد میں جلوس جمخانہ کالونی کی جانب روانہ
ہوگیا۔ جلوس کے شرکاء معزول چیف جسٹس افتخار محمد چودھری سے ملاقات کرنا چاہتے تھے۔ جمخانہ کالونی کے باہر
انہیں پولیس کی جانب سے شدید مزاحمت کا سامنا کرنا پڑا۔ میریٹ ہوٹل کے سامنے پولیس نے وکلاء کو پانی کے
پریشر سے منتشر کرنے کی کوشش کی۔ بعد میں پولیس نے آنسو گیس کے شیل پھینکے اور لاٹھی چارج کیا۔ دوسری
جانب وکلاء نے پولیس پر پتھراؤ کیا۔ شیلنگ اور پتھراؤ کا یہ سلسلہ پینتالیس منٹ تک جاری رہا اور ربڑ کی گولیاں برسائی

بسم الله الرحمن الرحيم
شروع کرتا ہوں اللہ کے نام
سے جو نہایت مہربان اور کریم ہے
یہ اردو تحریر کی ایک مثال ہے۔
امید ہے کہ اس کیبورڈ کے انداز سے
آپ سیکھو اردو لکھنے میں آسانی ہوگی۔
برائے کرم اپنے خیالات سے مجھ کو
آگاہ فرمائیے گا۔

Showing Text Line Extraction



انفارمیشن ٹیکنالوجی کے اداروں، پیشہ ورانہ تربیتی تنظیموں اور پروگرام مینجمنٹ کمپنیوں سے قومی سطح پر
پاکستانی طلباء کے درمیان مقامی اور مقامی زبانوں میں انٹرنیٹ پر معلومات کی تخلیق کے مقابلے کا غاز کرنا،
عملی شکل دینے اور انعقاد کے لیے پیش کشیں مطلوب ہیں۔
ہماری اس کاوش سے نوجوان نسل سائبر سپیس کی مربوط ترقی کے مواقع سے متعارف ہوگی۔
نیٹ
پا

Text Area Extraction

After text lines are identified the next step is to omit the extra white space in between lines and side ways which are of no use. The area which is covering the text is called text area. The algorithm takes pixel from the starting vertical point of the extracted text line to the next starting vertical point. Then examine the pixels first horizontally then vertically to find extra white spaces.

Showing Text Area Extraction

Compound Component Extraction

The whole compound ligature along with its primary and secondary stroke is called compound component. The idea behind the compound component extraction is to work on components to identify primary and secondary strokes instead of working with the whole text line. We are differentiating component with ligature because of overlapping in Urdu writing compound component can have more than one ligature or can have ligature with an isolated character. Algorithm first extract connected component starting from right to left. Each connected component is than inspected and compound component is formed according to the following rules:

If area of a connected component reside over another connected component than both are the parts of one compound component.

If area of a connected component reside under another connected component than both are the part of one compound component.

If area of a connected component reside over or under another connected component more than 50% than it is a part of one compound component.

If area of a connected component does not reside over or under another connected component than component it self is compound connected component

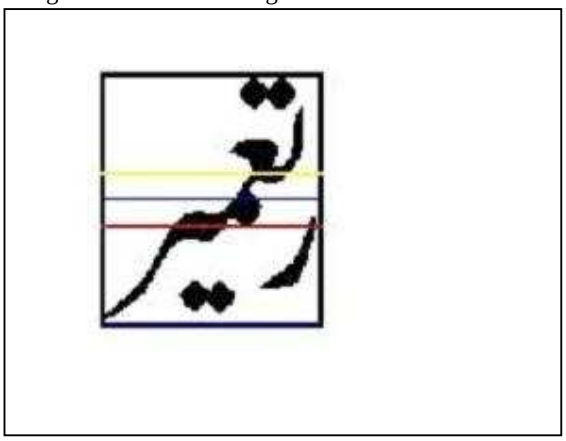
Showing Text Compound Component Extraction

Base Line Detection

Base line detection algorithm have mostly been applied on text line to identify primary and secondary strokes but in Urdu scripts the base line identification algorithms not showing the perfect result as it shows in Arabic scripts. To reduce the complexity we have applied the baseline algorithm on compound component. Base line is the horizontal point where maximum black pixels are present. Connected component that lies on the base line are primary

strokes and others are secondary strokes. For more accuracy we have sketch two horizontal lines on the basis of compound component statistics. These lines are called average horizontal lines. Lines are drawn on 50% and 35% of the height of compound component. These figures are calculated on the basis of average height of different compound component. These computations are forward to next stage that is stroke extraction.

Showing Base line and Average horizontal lines



Stroke Identification

Primary and secondary strokes are identified on the basis of base line and average horizontal lines according to the following rules:

Strokes that lie on base line and one of the average horizontal lines are primary strokes.

Strokes that lie on both average horizontal lines are primary strokes.

Strokes that do not lie on any line are secondary strokes.

Strokes that lie only on one average horizontal line are secondary strokes

If one stroke lies on base line and other stroke do not lie on any line than base line stroke is primary and other is secondary.

Results on the basis of above rules are quite accurate but for more accuracy especially for handwriting we have developed stroke identification improvement algorithm.

Improve Stroke Identification: More than one primary stroke can exist in one compound component and in handwriting it may identified as secondary stroke by above rules. For this we have developed a data on the statistics of text area of different hand writing and scripts. We have stored the ratio between text areas which is computed in text area extraction stage with the area of primary and secondary stroke. This data give us the average ratio between text area and area of primary and secondary stroke. These ratios are applied on the pre extracted stokes according to the following rules:

If any stroke which is marked as primary have ratio within the pre computed ratios of secondary than is marked as secondary.

If any stroke which is marked as secondary have ratio within the pre computed ratios of primary than is marked as primary.

If a stroke have ratio which is not in between the ratios of pre computed primary and secondary ration than it is not changed.

Features Extraction

We have computed five features for single character or ligature. We have extracted four features by sliding window technique and one feature is extracted by Hu Moment algorithm. First of all image is resize into 64x64 pixels after that features are extracted.

Sliding Window

We have take three types of sliding window which are:

8x64 pixles: 8x64 pixel window move from right to left and compute the ratio between white and black pixels.

*64x8 pixels:*64x8 pixels window move from top to bottom and compute the ratio between with and black pixels.

*c) 8x8 pixels:*8x8 pixels window moves from top right to bottom left and compute the ration between white and black pixels.

Square shape: Square shape method first read 2x2 pixels from right top to bottom left and if any black pixel is found than whole 2x2 pixel are converted in black pixels. This method gives an erect shape to charater converting loops in to perfect loops. The reason for this methodology is that sliding window technique gives better result on characer whos shapes are more straight. Than 8x64 pixels window move form left to right to compute ratio between white and black pixels.

Hu Moment

Hu invariant moments are calculated for character and ligatures.

Recognition

Recognition is the last step in order to achieve our desired output, the extracted features of character or ligature is then match with the stored features to recognize the character. We have used K-Nearest Neighbors (KNN) algorithm for features matching. In KNN we have applied Euclidean distance with 10 nearest neighbors. We have matched five features independently by KNN, the maximum same result given by independent result is the final recognition result.

EXPERIMENTAL RESULTS

We have developed the proposed OCR system on MATLAB and Microsoft C#.Net. Both online (*only for isolated characters)* and offline OCR were developed. The

system was tested on different printed and handwritten documents of different founts and script.

Our system gives 97.09% accuracy in extracting text lines. Accuracy of 98.86% found in primary and secondary stroke extraction. Recognition gives accuracy of 97.12%. The overall results of the Urdu OCR system was quite encouraging for both online and offline OCR.

Offline OCR

Online OCR

CONCLUSION

We have proposed a system that describes methodology for both online and offline Urdu OCR system. Although many researches have been done before but they were script specific.

Results of our developed system are promising but definitely more perfection is required.

FUTURE WORK

Future work includes improvement of algorithms. Complete document analyses needs to be done. Segmentation free approach required too much computation so research has to be done on character segmentation.

References

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

Inam Shamsher et.al, OCR For Printed Urdu Script Using Feed Forward Neural Network, Proceedings of World Academy of Science,Engineering and Technology. Vol 23, Aug 2007 ISSN1307-6884

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. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, p. 271–350.

Khalid Saeed, "New Approaches for Cursive Languages Recognition: Machine and Hand

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.]

U. Pal and A. Sarkar. Recognition of printed Urdu script. In SeventhInt. Conf. on Document Analysis and Recognition, pages 1183–1187,Edinburgh, UK, Aug. 2003.

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

S. Mozaffari, K. Faez, and M. Ziaratban. (2005). Structural Decomposition and StatisticalDescription of Farsi/Arabic Handwritten Numeric Characters. Proc. Int'lConf.Document Analysis and Recognition,pp. 237- 241.

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

Syed. Afaq Husain and Syed. Hassan Amin A Multi-tier Holistic approach for Urdu Nastaliq Recognition

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)

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