American Journal of Neural Networks and Applications
2020; 6(1): 16-21
http://www.sciencepublishinggroup.com/j/ajnna
doi: 10.11648/j.ajnna.20200602.11
ISSN: 2469-7400 (Print); ISSN: 2469-7419 (Online)

Urdu Nastaleeq Nib Calligraphy Pattern Recognition

Mateen Ahmed Abbasi[1], **Naila Fareen**[2], **Adnan Ahmed Abbasi**[3]

[1]Engineering and Information Technology, Khwaja Fareed University, Rahimyar Khan, Pakistan

[2]Department of Computer Science, Allama Iqbal Open University, Islamabad, Pakistan

[3]Department of Management Science, Alhamd Islamic University, Islamabad, Pakistan

Email address:
mateenabbasi@msn.com (M. A. Abbasi), nailafareen@hotmail.com (N. Fareen), adnanabbasi1706@yhoo.com (A. A. Abbasi)

To cite this article:
Mateen Ahmed Abbasi, Naila Fareen, Adnan Ahmed Abbasi. Urdu Nastaleeq Nib Calligraphy Pattern Recognition. *American Journal of Neural Networks and Applications*. Vol. 6, No. 2, 2020, pp. 16-21. doi: 10.11648/j.ajnna.20200602.11

Received: March 31, 2020; **Accepted**: April 10, 2020; **Published**: August 27, 2020

Abstract: Nib calligraphy pattern recognition is the way to convert handwritten nib font into its equivalent machine understandable or readable form. Nib calligraphy pattern recognition is derived from pattern recognition and computer vision, a variety of work has been done on Urdu literature and on Urdu handwritten automatic line segmentation. This research work is based on Urdu Nastaleeq Nib calligraphy pattern recognition. The width of the Qalam (Nib) makes difficulties in recognition due to different width of qalam pattern varieties, so there is dire need to develop a system that can recognize the digitized image of Urdu Nastaleeq Nib font with high accuracy. The objective of this research is to create a ground for the development of an efficient and robust Urdu Optical Character Recognition (OCR) for Urdu Nastaleeq nib pattern recognition and to develop a system that can recognize the digitized image of Urdu Nastaleeq Nib font with high accuracy. Urdu Nastaleeq nib pattern recognition. The research work mainly focuses on identifying the Urdu nib calligraphy pattern recognition. The purpose of the research is to create a system for Urdu Nastaleeq Nib calligraphy pattern recognition to get benefit from the cultural heritage of Nib calligraphic material. The Urdu Nastaleeq Nib Calligraphy Pattern Recognition research work is proposed to be done on the calligraphic Urdu Nastaleeq Nib pattern recognition. This research mainly focuses on recognizing the handwritten Urdu Nastaleeq Nib typeset and eliminating the noise which is the main difficulty in interpretation the font clearly. The aim here is to build up a more consistent, correct and precise system for Urdu Nastaleeq Nib calligraphy Pattern Recognition.

Keywords: Nib Calligraphy, Optical Character Recognition, Pattern Recognition, Urdu Nastaleeq

1. Introduction

The process of converting handwritten and typed characters into machine readable form proceeds in OCR (Optical Character Recognition). Nastaleeq nib calligraphy has eight nib points, these nib points were used in newspapers and on books which are written by hand in early seventies. Urdu Nastaleeq nib patterns are cursive in nature and are difficult to recognize [1]. Various methods for recognition of English, Latin, Chinese, Urdu and Arabic script have been proposed on hand written as well as on printed documents. The main problem in writing Urdu language is its different shapes due to attachment with other words at different position. Final and middle character attached with other character for a word may be different [2]. The Arabic writing system based on thirty alphabets which are easy to recognize but difficult to understandable when they are written by hand. These letters were categorized by supervised neural network. The writing system in that research was constituted with the help of mouse. The neural network checks the possibility of recognition in analog and it takes more than necessary time for training. The outcomes have shown good recognition rate for both distinct and continuous script. The Result considered best presented by the system for distinct script than of continuous [3]. The characters of Naskh script and related font are used by a larger quantity of the world’s population to write verbal communication such as Arabic, Persian and Urdu. OCR works by scanning documents and performing character analysis on the resultant figure [4]. Nastaleeq is a multifaceted font because of its nature. They present the clarification that uses Omega as the Typesetting Engine for rendering Nastaleeq. As describe that more than twenty thousand ligatures exist in Urdu script and used only seven thousand [5]. Urdu is written in different styles, shapes and it is spoken by more than sixty million speakers. There is a lot of work has been done on Urdu offline, online, typed and handwritten characters but no work has been done on Urdu Nastaleeq Nib pattern recognition so there is dire need of recognizing the nib calligraphy patterns in order to get benefit of the cultural heritage written with Nib.

2. Related Work

Urdu Nastaleeq nib font is bit different from other fonts used for Urdu. In Urdu, each character has two to four different shapes depending upon its position in the word: isolated, initial, medial or final. Figure 1 shows four different shapes of Urdu letters Nastaleeq nib calligraphy patterns.

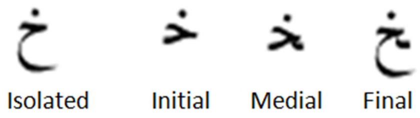


Figure 1. *Compound characters.*

Urdu Nastaleeq nib calligraphy practice sheet in figure 3 shows the different characters below and above the base line.

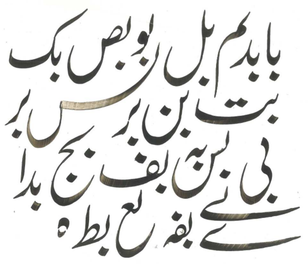


Figure 2. *Urdu Nastaleeq nib calligraphy practice-sheet.*

The recognition of the word is complex when a letter comes at the start, in the middle or at the end it changes its basic shape [6]. In the last forty decades OCR is one of the most significant fields of pattern recognition [7]. Urdu OCR Naskh typeset pattern matching methods worked on isolated characters of Naskh font for Unicode based computer [8]. Self organizing map (SOM) is used for classification of different characters with different shapes. is also being use for grouping of different characters used self organizing map to classify the Naskh characters into thirty three classes by automatic grouping of same ligature for early classification [9].

Feed forward neural network used for the categorization of Nastaleeq compound characters the model was created in matlab and it achieved 70% correctness on average. Strong techniques are required for detection of Urdu compound characters. Neural networks are self-possessed of simple inputs and outputs nodes use in comparable method [10]. Trouble-free Optical Character recognition (OCR) method that can be built without using neural network and can be built by using ''softconverter''. This paper verifies a model of softconverter and shows accuracy of 97% on average rate [11]. Urdu Nastaleeq is combination of Arabic and Persian language, it forms from Naskh and Taleeq. Urdu holds some properties of Arabic language and some properties of Persian language [12]. Nastaleeq is originated from Taleeq which is originated from Naskh. They use support vector machine for sorting of Naskh and Nastaleeq font. As Urdu is originated from Arabic writing script like Naskh font it starts from right [13]. Optical character recognition is classified into main two categories online and offline. Urdu script is categorized in mainly four shapes these are isolated shape. End shape, start shape and middle shape. Urdu writing style is used in widely in Asia and in Arab countries. Nastaleeq has more single characters than of Naskh and Taleeq writing script [14].

Much work has been done on Urdu handwritten sentence database that line segmentation of Urdu handwritten sentence database were correctly segmented according to the comparison only 2.4% was not rightly divided [15]. Word spotting in Gray-scale Pashto Documents, written in modified Arabic scripts. The approach has effectively handled the handwriting variations of 200 different writers. The average precision rate achieved is 94.75% for an average recall of 60.25% [16]. The authors focuses CPU- GPGPU combination using CUDA platform for software development of SOM algorithm. The images of different N x N dimensions are feed as input to the SOM network and image clustering is achieved through SOM training in the form of final weight matrix [17]. The problem of holistic recognition of printed ligatures in Nastaleeq writing style of the Urdu language, the main difficulty of the recognition process lies in the large number of classes/ligatures (17,000 different possible ligatures in our Urdu text data). They presented the idea to build the foundations for practical large-scale ligature classification systems not only for Nastaleeq, but also for other Urdu and Arabic scripts [18].

3. Nastaleeq Nib Calligraphy Font

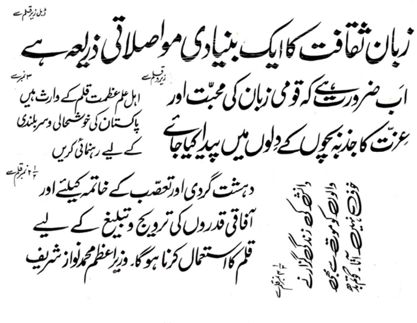

زبان ثقافت کا ایک بنیادی مواصلاتی ذریعہ ہے
اب ضرورت ہے کہ قومی زبان کی محبت اور
عزت کا جذبہ بچوں کے دلوں میں پیدا کیا جائے
دہشت گردی اور تعصب کے خاتمہ کیلئے اور
آفاقی قدروں کی ترویج و تبلیغ کے لیے
قلم کا استعمال کرنا ہوگا۔ وزیراعظم محمد نواز شریف

Figure 3. *Different Qalam Size Image.*

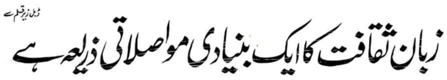
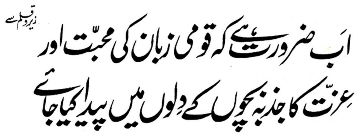


Figure 4. *Double zero Qalam Size.*

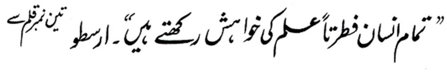
These words are taken from the sample. Its Nib number is double zero.



زیرو قلم سے
اب ضرورت ہے کہ قومی زبان کی محبّت اور
عزّت کا جذبہ بچوں کے دلوں میں پیدا کیا جائے

Figure 5. *Zero Qalam Size.*

Words written with Nib zero are greater in size from Nib double zero.



" تمام انسان فطرتاً علم کی خواہش رکھتے ہیں۔ ارسطو
تین نمبر قلم سے

Figure 6. *Qalam Size Three.*

These words are taken from the sample and its Nib size is three. The nib size increases the word size decreases. Similarly when the nib size increases to eight the word size will become smaller.

4. Methodology

The research method to be pursued that Urdu Nastaleeq Nib Calligraphy patterns taken from scanning devices will be used as input image, after collection of images the training process will start and trained the images after downsampling the recognition process will start to make the Urdu Nastaleeq nib calligraphic font in editable form.

A. Knowledge Base

In the proposed system knowledge base is involved from the acquisition of the image to image recognition. Figure 6 shows the complete knowledge base of the proposed system.

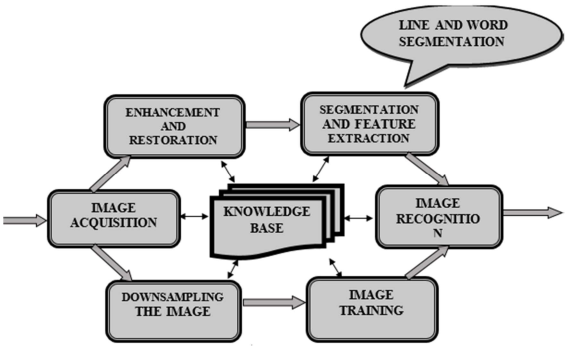

LINE AND WORD SEGMENTATION
ENHANCEMENT AND RESTORATION
SEGMENTATION AND FEATURE EXTRACTION
IMAGE ACQUISITION
KNOWLEDGE BASE
IMAGE RECOGNITIO N
DOWNSAMPLING THE IMAGE
IMAGE TRAINING

Figure 7. *Knowledge Base.*

B. Image Acquisition

Image acquisition is the basic step for the recognition of a pattern or image [19]. The images are being acquired by camera, digital capturing devices and scanners. For the preprocessing phase image scanning method is used.

C. Image Enhancement

Image enhancement is consisting of techniques that convert an image into a shape matched for analysis to search for improving the visual form of the image. Image enhancement technique is used to highlighting the concealed details of curiosity of an image.

D. Binarization

The process of converting gray scale image into black and white (binary) image is called Binarization. The black pixel denotes the text and white pixels represents the background. [20]. Binarization plays an important role in digital image processing. In document processing for recognizing text and symbols binarization is used. Binarization assigns background or foreground to each pixel and determine the gray threshold [21].

5. Feature Extraction

Feature extraction extract properties to identify a character uniquely and differentiate between similar characters. Easily and correctly recognized character by human can be written in variety of ways. Data collection methods, feature extraction principles and classification methods are used in the field of character recognition [22].

6. Image Thresholding

The classification of background and foreground pixels of an image is called image thresholding. The general idea of thresholding is convert pixels below the certain level of gray into background and to convert pixels above the certain level of gray into foreground and the value of gray scale image is from 0-255. White pixels show the background while the foreground is represented by black pixels.

7. Normalization

To obtain standardized data all the variations during the writing is removed by normalization method. Normalization can be skew normalization, slant normalization and size normalization.

8. Holistic Method

Holistic method is used for the Urdu Nastaleeq Nib Calligraphic word recognition. In handwriting style the words are generally complex patterns and comprise excessive inconsistency. Hand written word recognition significantly aided by a vocabulary of valid words reliant on the application area. The segmentation based approach is called the analytical approach and the segmentation free or word based is called holistic approach. In holistic approach the glyph, ligature and words can be recognize as a whole, there is no need to divide them in subunits. The holistic paradigm was inspired by psychological studies of human reading by ascenders, descenders and length of the word used by human as features of word shape. [19].

9. Self Organization Map

Self Organization Map (SOM) is used for character recognition of Urdu Nastaleeq Nib font. Neural network is used for the input neurons and output neurons. SOM is architecture of neural network. SOM's are often used to group input and are usually trained with an unsupervised training algorithm. An SOM uses a winner-takes-all strategy, in which the output is provided by the winning neuron. SOM is the core module of this research work. It is the topology preserving vector quantization algorithm. Framework of neurons (‘nodes’) accepts and responds to set of input signals. Framework compared; selected ‘winning’ neuron. Selected neuron activated together with ‘neighbourhood’ neurons. Similar input changes weights through adaptive process. The software can conclude the winning neuron by matching the values of the output neuron. Winning neuron is the biggest output value.

A. Why SOM

In supervised neural networks the desired response of the system is already known. In supervised training, the iteration proceeds until the output of the given sample matches with the expected output. While in Self SOM the output is not known. In unsupervised training the output trained to answer to group of patterns with the input. In this paradigm, the system is supposed to discover statistically salient features of the input population. [22].

Topology preserving is referred as Self Organization Map (SOM) for the reason that it preserves the comparative distance between points. SOM is just having input and output nodes excluding hidden nodes. SOM produce only single value either 0 or 1. As a result, the output from the SOM is usually the index of the fired neuron. [23].

10. Segmentation

Segmentation is the method to breakdown an image or text into subparts for detection and recognition. Because of complex nature of Urdu script segmentation for Urdu text become difficult [24]. The main problem in writing Urdu language is its different shapes due to attachment with other words at different position. Final and middle character attached with other character for a word may be different.

A. Line Segmentation Algorithm

The line segmentation algorithm represents the segmentation between the lines. If the system input is a full page text, the algorithm helps in separating the lines of the paragraph and saves each line in the output directory.

1. Input text image.
2. Create an output directory.
3. Width, height and text format should be according to the predefined format.
4. Begin scan from right peak of the original text image.
5. Scan horizontally; consider line width-wise and find the top to bottom of each vertical line.
6. “Black” pixels checked by the scanning of first row. If no “black” pixel is found move to next row.
7. Keep on scanning until rows containing “black” pixels is come across.
8. Stop scanning when another row with all “white” pixels is found. Set this row as the bottom edge of the image text line.
9. Save image as first text line numbered each line, continue scan to find next line and save the segmented lines in the output directory.

B. Word Segmentation Algorithm

Word segmentation algorithm presents the segmentation approach for words of the OCR system. Each word is separated by the space between the words.

1. Load text line image.
2. Create an output directory.
3. Scan up to the image height, on the horizontal projection.
4. “Black” pixels checked by the scanning of first row. If no “black” pixel is found move to next row.
5. Keep on scanning until rows containing “black” pixels come across.
6. Stop scanning when another row with all “white” pixels is found. Set this row as the bottom edge of the image text line.
7. Save image as first word, numbered each word and continue scan to find next word.
8. Save all word images in the output directory.

11. Experimentation

The outcome can be observed in three perspectives; paragraph identification rate, sentence identification rate and character identification rate. The system produces 100% result on the recognition of basic isolated characters of Urdu Nastaleeq Nib Calligraphy font, Urdu digits written with Nib. The system is also tested on the symbols written with Nib. It produces 89.03% results of the overall recognition. Basic characters, numbers, symbols, recognized character ligatures, recognized images and their accuracy % age are mentioned in the Table 1.

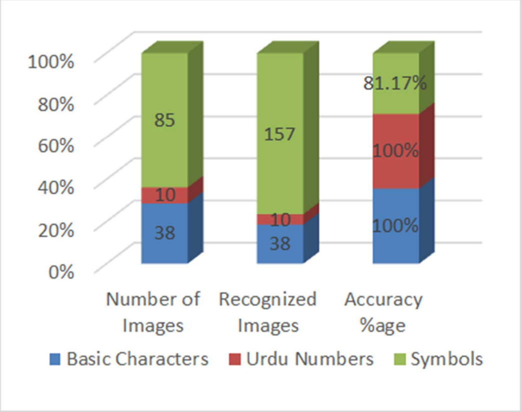


Figure 8. *Basic characters.*

Table 1. *Result of Total images, Recognized images and Accuracy % age.*

Image	Number of Images	Recognized Images	Accuracy % age
Basic Characters	38	38	100%
Urdu Numbers	10	10	100%
Symbols	85	69	81.17%
2 character ligature	627	502	74.96%
Total	760	587	89.03%

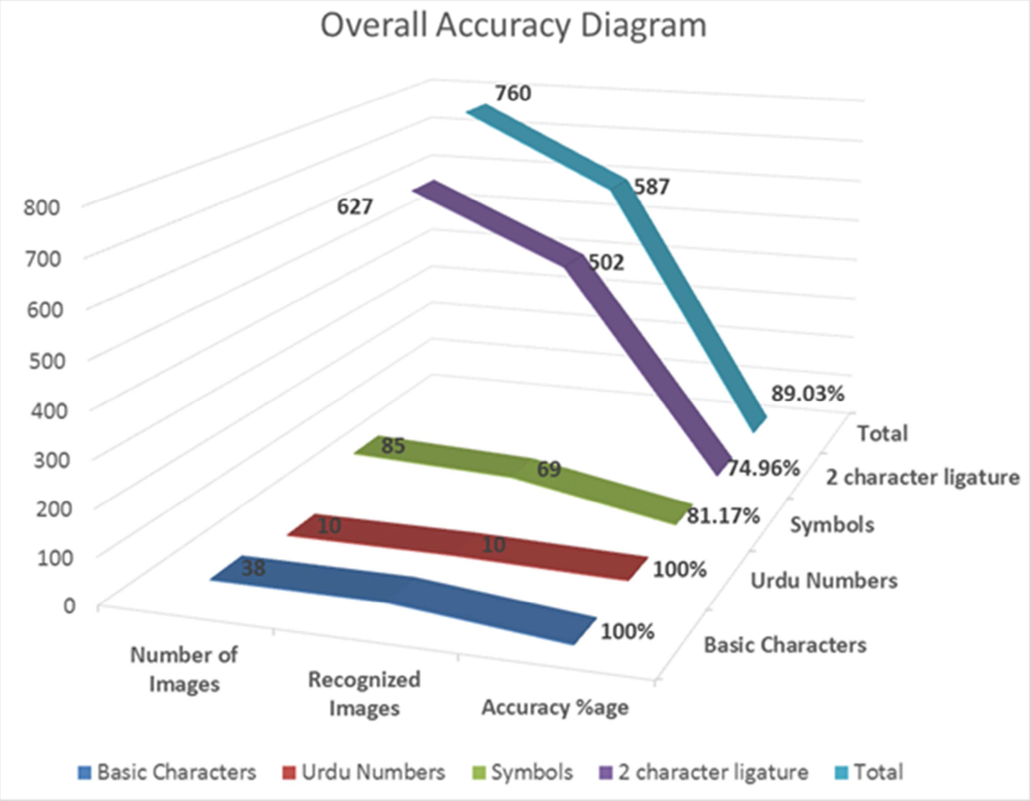


Figure 9. *Overall accuracy % age.*

12. Conclusion

The Research is based upon Urdu Nastaleeq Nib Calligraphy Patter Recognition, Urdu Nastaleeq Nib Calligraphic Pattern are being taken from the corpus of National Language Authority Islamabad Pakistan All the material is written with the Nib of different sizes from point double zero to point eight. Literature survey of Urdu Farsi and Arabic OCR are discussed. In this research neural network technique SOM is used for the recognition procedure. The research shows the complete recognition process of Urdu Nastaleeq Nib Patterns, algorithm made for recognition process, line segmentation and character segmentation process. In this research the prototype is cable of recognizing the basic characters Urdu digits from zero to nine, symbols and two ligature characters and produced 89.03% recognition rate. The selected research work is about the OCR development for Nib Calligraphy font. The proposed work's aim is to make rules and develop OCR which recognizes the Nastaleeq Nib font and convert it into machine readable form.

References

[1] Naz. s, Arif I. Umar R, Ahmad S. B. Ahmed, S. H. Shirazi, M. I. Razzak. 2017. Urdu Nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features, Neural Computing and Applications, V. 28, pp 219-231.

[2] Hasan, U. A., S. B. Ahmed, S. F. Rashid, F. Shafait and T. M. Breuel. 2013. Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks, 12th International Conference on Document Analysis and Recognition, pp. 1061-1065.

[3] Sheikh, S. 2010. Arabic-Urdu script Recognition through Mouse: An Implementation using artificial neural Network, Seventh International Conference on Information Technology, pp. 307-310.

[4] Steven M. B., Eric C. J. & David A. G. Retrieving OCR Text: A Survey of Current Approaches, 2003. ACM SIGIR, 36 (2), pp 58-61.

[5] Atif G. &Shafiq R. Nastaleeq: a Challenge Accepted by Omega, Tugboat, 2007. XVII European TEX Conference, 29 (1), pp 89-94.

[6] Zaheer A, J. k. Orakzai., I. Shamsher and A. Awais. 2007. Urdu Nastaleeq Optical Character Recognition, World Academy of Science, Engineering and Technology, pp. 249-252.

[7] Haidar A., John S. G. &Hisham A. A Real-time DSP-Based Optical Character Recognition System for Isolated Arabic Character using the TI TMS320C6416T, 2008. Proceedings of the 2008 IAJC-IJME International Conference.

[8] Tabassam, N., S. A. H. S. Naqvi, H. Rehman and F. Anoshia. 2009. Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique, International Journal of Image Processing, vol. 3, no. 3, pp. 99-104.

[9] Hussain, S. A., S. Zaman and M. Ayub. 2009. A Self Organizing Map Based Urdu Nasakh Character Recognition, International Conference on Emerging Technologies (ICET), 19-20 Oct Islamabad, Pakistan, pp 267–273.

[10] Zaheer A, Jehanzeb k. O., Inam S. IEEE International Conference on Computer Science and Information Technology, 2009. ICCSIT 2009 8-11 Aug. Beijing, China. pp 457–462.

[11] Junaid T, Umar N, Muhammad U. N. Softconverter: A novel approach to construct OCR for printed Urdu isolated characters 2010. 2nd International Conference on Computer Engineering and Technology (ICCET), 2010 16-18 April. Chengdu, China. pp V3-495 - V3-498.

[12] Shuwair S, Abdul W. 2010. Optical character recognition system for Urdu International Conference onInformation and Emerging Technologies (ICIET), 2010 14-16 June Karachi, Pakistan pp 1-5.

[13] Sagheer, M. W., C. L. He., N. Nobile and C. Y. Suen. 2010. Holistic Urdu Handwritten Word Recognition Using Support Vector Machine, Proceedings of 20th International Conference on Pattern Recognition (ICPR), pp 1900-1903.

[14] Bukhari, S. S. and T. M. Breuel. 2011. Generic Layout Analysis of Diverse Collection of Documents, International Conference on Document Analysis and Recognition (ICDAR), pp 1275–1279.

[15] Ahsen R, Imran S, Ali A, Fahim A. 2012. An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation, International Conference on Frontiers in Handwriting Recognition (ICFHR), 201218-20 Sept Bari, Italy, pp 491–496.

[16] M. I. Shah, J. Sadri, C. Y. Suen, and N. Nobile, "A New Multipurpose Comprehensive Database for Handwritten Dari Recognition," Eleventh International Conference on Frontiers in Handwriting Recognition, Montreal, Canada, August 2008 pp. 635-640.

[17] S. Qasim, M. A Ismail, “Design and Implementation of Parallel SOM Model on GPGPU”, 5th International conference on computer science and information technology IEEE CSIT 2013, March 28-29, Amman, Jordan, pp 233–237.

[18] Akram El-Korashy, Faisal Shafait Search space reduction for holistic ligature recognition in Urdu Nastaleeq script, 2013, ICDAR, 12th International Conference on Document Analysis and Recognition, pp 1125–1129.

[19] Cheriet, M. N. Kharma, C. L. Liu and C. Y. Suen. 2007. Character Recognition Systems A Guide For Students And Practioners, Published By John Wiley & Sons.

[20] S. Vavilis and E. Kavallieratou. 2011. A tool for tuning binarization techniques, International Conference on Document Analysis and Recognition, (ICDAR), pp. 1-5.

[21] Yousefi, J. 2011. Image Binarization using Otsu Thresholding Algorithm, Image processing and digital image processing, International workshop on Document Analysis Systems, pp. 1-4.

[22] Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen. 1996. SOM PAK: The Self-Organizing Map program package, pp. 1-27.

[23] Heaton, J. 2008. Introduction to Neural Networks for C#, Heaton Research, ISBN 1-60439-009-3.

[24] Naz, S., K. Hayat, M. I. Razzak, M. W. Anwar and H. Akbar. 2013. Arabic script based character segmentation: A review, World Congress on Computer and Information Technology (WCCIT), pp. 1-6.