IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Deep Learning Techniques for Naskh and Nastalique Writing Style Text Recognition

Shanky Goel[1*], Gurpreet Singh Lehal[2]

[1,2]Department of Computer Science, Punjabi University, Patiala, India

[*]*Corresponding Author: sgoel9803415203@gmail.com, Tel.: 9803415203*

DOI: https://doi.org/10.26438/ijcse/v7i4.7076 | Available online at: www.ijcseonline.org

Accepted: 17/Apr/2019, Published: 30/Apr/2019

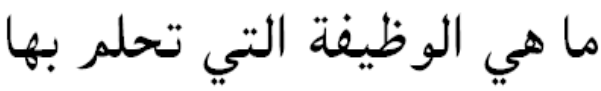
***Abstract*—** Naskh and Nastalique text recognition are a challenging task in the Pattern Recognition field because of the cursive and context sensitive nature of the script. Many languages use Naskh or/and Nastalique style for writing. Due to the complexities associated with these writing styles, not much effort has been done for the development of real-time recognition systems for Naskh and Nastalique writing style languages. Traditional recognition process segments the text image into characters for subsequent OCR phases which is less accurate for Naskh/Nastalique text and reduces the accuracy of the recognition system. Recently, Recurrent Neural Network (RNN) based Long Short Term Memory (LSTM) architecture with Connectionist Temporal Classification (CTC) has shown a remarkable result in text image recognition. This paper presents the recognition challenges in the Naskh and Nastalique writing style text and a study of different deep learning techniques applied for the recognition of Naskh Arabic and Nastalique Urdu text.

***Keywords*—** Naskh, Nastalique, Recognition Challenges, RNN, LSTM.

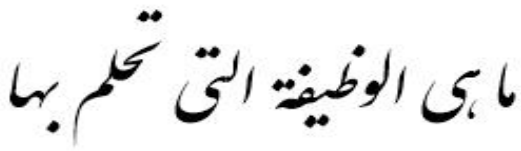
I. INTRODUCTION

Many languages such as Sindhi, Persian, Pashto, Kurdish etc. are based on Arabic alphabet and follows Naskh style of writing. In Naskh writing style, characters are connected on an imaginary/horizontal line called baseline. Arabic alphabet based languages follows the characteristics of Arabic script and differs in number of characters only. Despite the decades of the research in the field of Arabic OCR, not much effort has been done for the recognition of Arabic alphabet based languages. There is a lack of real time OCR systems for Arabic script adopted languages. The main reason behind this is the late start of the research in Arabic OCR and complexities associated with the script. Many Arabic alphabet based languages use Naskh writing style. Nastalique is more stylistic and calligraphic as compared to Naskh writing style. From OCR point of view, Nastalique is more complex than Naskh. Nastalique style is widely used for writing Kashmiri, Punjabi and Urdu languages. It is also used alongside with Naskh for Pashto and only for poetry in Persian. Urdu OCR problem is much similar to Arabic because Urdu uses an extended and adapted Arabic script. Urdu has a total of 39 characters whereas Arabic has 28 characters. Urdu has mainly two writing styles: Naskh and Nastalique (Figure 1). Nastalique is usually used for writing Urdu language. Development of OCR system for Urdu language is a challenging task due to the Nastalique writing style of the script. In this paper, our aim is to present the recognition challenges in Naskh and Nastalique writing style text and deep learning based network solutions for the recognition of these style text.

This paper is organized as follows. Section II describes the characteristics of Naskh and Nastalique writing style based languages. Section III presents the challenges in the recognition process of Naskh and Nastalique text. Section IV presents the deep learning techniques used for the recognition of Naskh and Nastalique text. Finally, a conclusion is presented in Section V.



(a)



(b)

Figure 1: (a) Naskh (b) Nastalique writing style text

II. NASKH AND NASTALIQUE TEXT CHARACTERSTICS

- Naskh and Nastalique based languages are cursive in nature. Machine printed Arabic script follows Naskh style in which characters are connected on an imaginary line. However, Nastalique is highly cursive as compare to Naskh and written as diagonally from top right to left with stacking of characters.
- Characters in text make no distinction between lower and upper case. It contains only one case.
- Text is written from right to left both in printed and handwritten forms, while numbers are written from left to right (Figure 2).

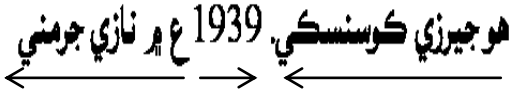


Figure 2: Bi-directional writing

- In Naskh and Nastalique languages, the shape of the character varies according to its position (initial, middle, final and isolated) in the word/sub-word. For example, Arabic character ع is written as عـ in initial, ـعـ in middle, ـع in final and ع in isolated positions. Character on different positions not only changes its shape but also its size. So, width and height of the character are inconstant.
- The character shapes join together to form words or ligatures. A ligature/sub-word is a part of a word or sometimes a complete word, in which all character shapes must be connected together. A word in Naskh/Nastalique text is composed of ligatures and isolated characters, for example the word كتاب is formed by one ligature and an isolated character.
  Below ligature is composed of 3 characters.
  كتا = ك+ت+ا
  Ligature Characters used in ligature
- Naskh and Nastalique script based languages are context sensitive i.e. characters in a ligature change shape depending upon their position and preceding or succeeding characters.
- Diacritics and dots are used in these writing style based languages (Fig.3). Dots can be placed above, inside or below the character. Diacritics can be marked above or below the character.

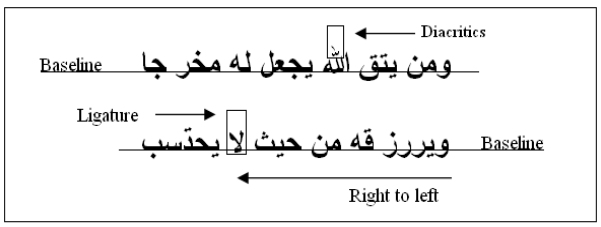


Figure 3: Characteristics of Arabic script

III. RECOGNITION CHALLENGES IN NASKH AND NASTALIQUE TEXT

The cursiveness and context sensitivity are the two major problems in the development of Arabic script based OCRs. Due to these problems, segmentation errors occur at the segmentation stage and it creates hindrance in the overall accuracy of the system. The main challenges in the development of Naskh and Nastalique script OCRs are:

A. Segmentation Challenge

Segmentation of a text image document into lines, words, and characters, is a crucial stage in OCR. The Arabic text image segmentation methods can be categorized into two approaches: Analytical Approach and Holistic Approach/Segmentation-Free Approach (Figure 4).

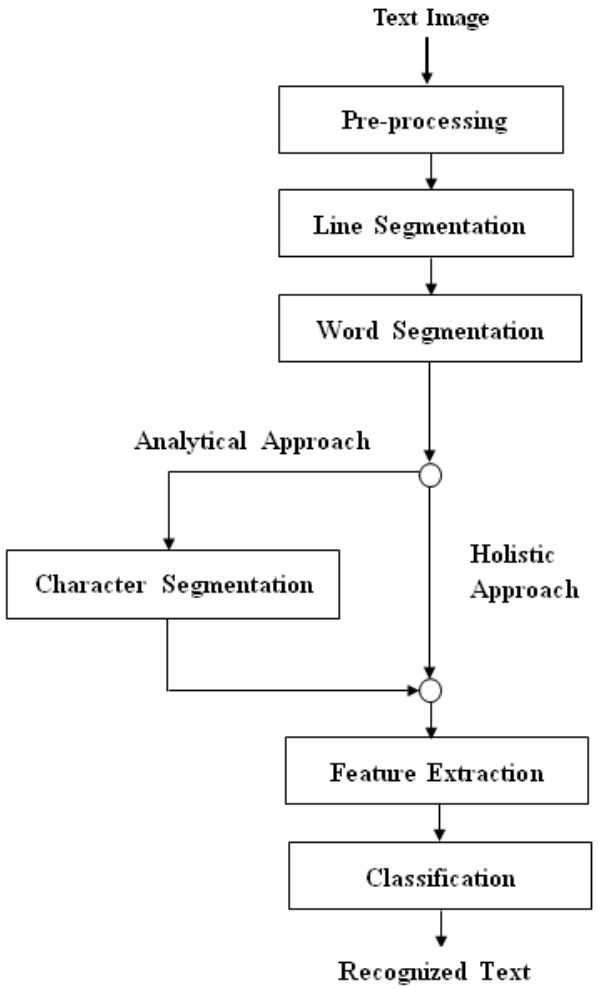


Figure 4: Text recognition approach for Naskh/Nastlique

1) Analytical Approach

In analytical approach, the document is segmented into text lines and then the text lines are segmented into the ligatures/words and further, ligatures/words are segmented into characters or primitives.

Characters are given individual recognition in this technique. Segmentation of a word into ligatures and isolated characters can be done by using connected component analysis. But character segmentation is one of the most challenging tasks, because false segmentation will result in misrecognition of the characters. Following are the main challenges in Analytical approach:

- It becomes most difficult task to find the accurate segmentation points, because of the cursiveness and context-sensitive nature of the script. The ligatures in a sentence have to be segmented into the individual character parts (Fig.5 and Fig.6), but it is very difficult to describe the correct cutting points. Character segmentation is almost impossible in Nastalique as compared to Naskh because in naskh segmentation points can be located at baseline but nastalique has no such line and written diagonally.

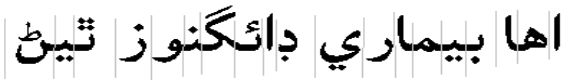


Figure 5: Character Segmentation of Naskh text line

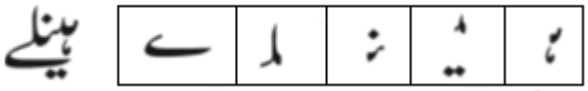


Figure 6: Character segmentation of Nastalique ligature [1]

- Character Overlapping is also a challenge in character segmentation, as it may result in over and under-segmentation of the character. Characters in a word may overlap even without touching each other, as shown in the figure below.

Figure 7: Character overlapping

- The character in a word may touch adjacent character or adjacent character's secondary components (dots/diacritics) which possess difficulty in character segmentation (Figure 8).

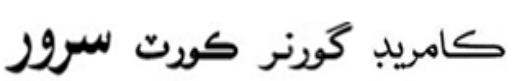


Figure 8: Touching adjacent characters in a word

In spite of these challenges in analytical approach, there are a large number of classes for recognition including joiner and non-joiner character's shape, diacritics, numerals, punctuation marks, and the minimum recognizable units/symbols, but it is really hard task to break text into these units, because of issues in the character segmentation stage.

2) Holistic Approach (Word or Ligature based approach)

Due to difficulty in character segmentation stage, usually holistic approach is used. In this technique, whole ligature or word is recognized without segmenting the word/ligature into characters. A large number of ligatures or words are identified, trained and recognized in the holistic approach. The exact number of ligatures/words to be trained for recognition is unknown in particular language. This is very time-consuming and leads to difficulty in identifying such a large number of ligatures/words. In the holistic approach, many times primary and secondary components are recognized separately. The primary component represents the basic shape of the ligature, while the secondary connected component corresponds to the dots and diacritics marks and special symbols associated with the ligature [1]. Sometimes, it becomes difficult to extract the secondary components from the primary component, as both are merged together; shown in the below figures (Fig.9 and Fig.10). Broken characters in ligature/word also complicate the recognition process (Fig.11).

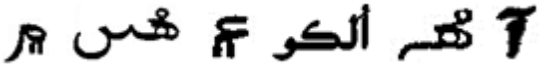


Figure 9: Diacritics merged with primary components

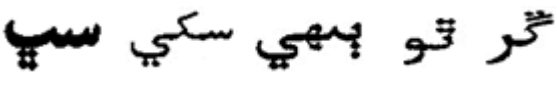


Figure 10: Dots merged with primary components

Figure 11: Broken characters in a ligature/word

B. Challenges in Dots

Arabic OCR's accuracy highly depends on correct recognition of dots. One noisy dot can change one character to another (Fig.12).

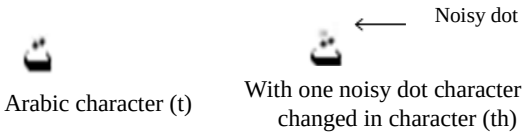


Figure 12: Noisy dot in character

Dots are an important part of characters in Naskh and Nastalique text as numbers; location and direction of the dots change the meaning of the character. Following situations of ambiguity may arise due to this. First, characters can have same base shape and same position of the dots, but may differ in number of dots (Fig.13).

Figure 13: Differ in the number of dots

In the second case, characters can have same base shape and an equal number of dots, but may differ in the position of the dots (Fig.14).

Figure 14: Differ in position of dots

In the last scenario, characters can have same base shape, the same number of dots, and same position of the dots, but may differ in direction of dots (Fig.15).

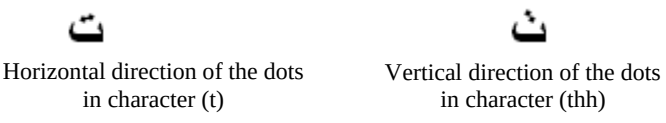


Figure 15: Differ in direction of dots

A dot may appear as merged or separated. It complicates the recognition process when dots and diacritics are merged together (Fig.16).

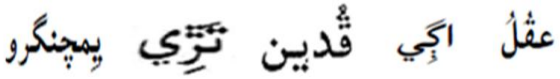


Figure 16: Merged dots and diacritics

Due to challenges in segmentation stage and correct recognition of dots, Recurrent Neural Network-Long Short Term Memory based architecture becomes current state of the art for cursive and Arabic script recognition.

IV. DEEP LEARNING

OCR research [2] is part of many other broad research areas like artificial intelligence, pattern recognition, machine learning, digital image processing, computer vision, and natural language processing. Deep learning sub-field of machine learning recently has gained more attention for image recognition. High computation based deep learning methods like RNN-LSTM, CNN, Auto-encoders, etc. achieved remarkable results in cursive text recognition. Due to high challenges in recognition process for Naskh and Nastalique writing style script, deep learning approaches are more suitable as it can process the text image as line or word without explicit character segmentation. LSTM based architectures have achieved a good recognition rate for Naskh and Nastalique text in spite of high challenges in OCR research for these writing style text. These architectures can be applied for other Naskh/Nastalique script languages. RNN-LSTM architectures are usually based on supervised learning, so text image corresponding with its transcription file both are fed to the model. Thus, a database of text images with its corresponding ground truth data is required for training the network. Following section describes the deep learning approaches applied for cursive text recognition.

A. RNN-BLSTM

Long-Short Term Memory (LSTM) based Recurrent Neural Networks (RNN) has the property of sequence learning and thus, suitable for the sequence problems such as text recognition, speech recognition, time series prediction problems etc. Bi-directional Long-Short Term Memory (BLSTM) a variant of RNN architecture scans the sequence in both forward and backward direction to capture the context. Ul-Hasan et al. [3] used deep neural network for the recognition of printed Nastalique Urdu text. They have applied BLSTM architecture with Connectionist Temporal Classification (CTC) layer on normalized text line images of printed Urdu. The text line images are taken from synthetically generated UPTI dataset. The system is evaluated for two cases: one ignoring the character's shape and the second is considering them. For the first case it achieved an error rate of 5.15% and 13.6% for the second case. Further, Ul-Hasan et al. [4] proposed LSTM architecture for multiple script recognition in normalized text-line images. They have developed a synthetically generated English and Greek text lines database and achieved a character recognition rate of 98.186% on the dataset. CTC layer [5] does the sequence alignment of input and output labels and thus does not require any pre-segmented data. Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Network (RNN) architecture with Connectionist Temporal Classification (CTC) layer has applied to evaluate its performance on Latin and Urdu scripts [6]. A synthetic database named 'Urdu-Jang Dataset' containing a total of 26,925 text lines has been developed. Raw pixel values are used as features; no other handcrafted features are extracted. The character recognition accuracy of 88.94% achieved for Nastalique Urdu text. Riaz Ahmad et al. [7] presented the recognition system based on deep learning for Naskh Pashto text line images. They have created the database of 17, 015 images having Pashto text lines. In BLSTM network, height of the input line image is normalized of 48 pixels. One hidden layer of 120 LSTM cells is used and character error rate of 16.16% has reported.

B. RNN-MDLSTM

Multi-dimensional Long-Short Term Memory (MDLSTM) based RNN architecture scans the image into four dimensions left, right, up, and down (Figure 17). Usually, MDLSTM network has applied on non-normalized images because it can handle the variation in vertical and horizontal direction of 2- dimensional images. Input block takes the raw pixel values from the input image and processed it in further layers. Hidden block takes the features from the MDLSTM layer and fed to the feed-forward layer for further processing. Graves introduced the MDLSTM system with CTC output layer [5] and reported the accuracy of 91.85% by using this system for Arabic characters [8]. They have trained the system by taking raw pixel values. In MDLSTM network, hierarchical subsampling window of size 1*4 is applied to input block and three hidden layers of BLSTM cell sizes 4, 20, 100 and two feed-forward tanh layers of sizes 16, 80 are used to recognize the Pashto text lines [7]. MDLSTM with CTC layer is also applied for the recognition of low-resolution images of Arabic text [9]. The authors used APTI database to train and evaluate the system and reported 99% word recognition rate for multi font, multi size Arabic text. The authors have trained MDLSTM network on 390 images and reported 9.22% character error rate. Ahmad et al. [10] evaluated the performance of different approaches include Long Short Term Memory (LSTM) network, Hidden Markov Model (HMM), and Scale Invariant Feature Transform (SIFT) for the recognition of Naskh Pashto cursive script. They have introduced a database of 480,000 ligatures/sub-words Pashto images. The database contains 1000 unique ligatures/sub-words and each unique ligature has 12 rotation and 40 scale variations. They have reported that performance of LSTM (98.9%) is better than HMM (89.9%) and SIFT (94.3%) for the recognition of Pashto. The LSTM based network has also applied for the recognition of Arabic handwritten dataset; known as KHATT [11]. MDLSTM model have trained with CTC output layer on Arabic handwritten text line images. The system improved the results from 46.13% to 75.8% for KHATT database. A dropout

layer is applied in MDLSTM-RNN architecture to recognize Arabic handwritten words [12]. The authors have used 6, 20 dropout layer size. The system has evaluated on IFN/ENIT database and reported the 12.09% label error rate. RNN model is also applied on Nasta'liq script by Naz et al. [13]. They have presented the MDLSTM based recognition system for Urdu text line images. They have applied hierarchical subsampling window of size 4*1 to input block and

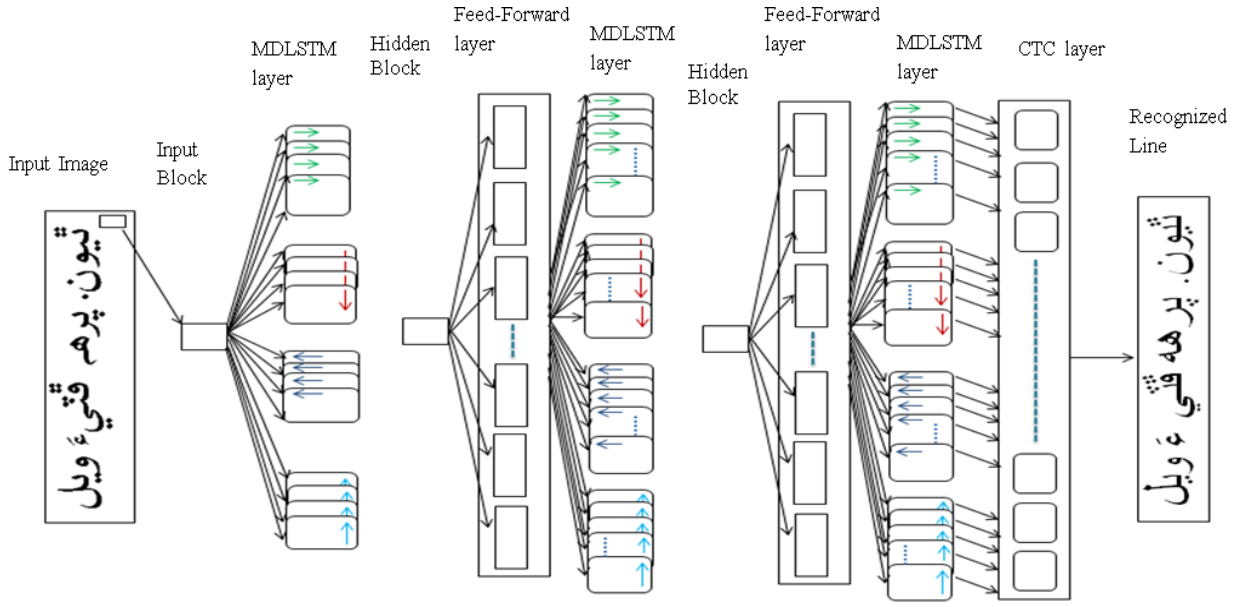


Figure 17: RNN-MDLSTM architecture

used three hidden layers which consist of BLSTM cells. The authors have evaluated their system on UPTI database and reported 98% accuracy. An implicit segmentation based recognition system for Nastalique Urdu text lines is also presented in [14]. The system extracted a set of statistical features by sliding overlapped windows on text line image. Further, these features are fed to RNN-MDLSTM with a CTC output layer. The authors evaluated the proposed technique on UPTI database. Naz et al. [15] used handcrafted features in MDLSTM network to recognize printed Urdu Nastalique font. They have extracted statistical features i.e. vertical edges intensities, horizontal edges intensities, density function, intensity features, foreground distribution, mean and variance of horizontal projections, mean and variance of vertical projections, center of gravity X and Y, and GLCM features by using a sliding window of size 4 * 48 (width * height) from right to left on normalized text line image. Further, these features are fed to MDLSTM network for recognition. The system achieved a recognition accuracy of 94.97% on UPTI database. Naz et al. [16] used zoning features and 2DLSTM learning classifier for the recognition of Nastalique Urdu text line images. The proposed model is evaluated on UPTI dataset with character recognition rate of 93.39%. Details of LSTM-RNN network parameters used to recognize Naskh and Nastalique wrting style languages are given in table 1.

C. CNN-RNN

A combination of Convolution Neural Networks (CNN) and MDLSTM approach has applied for the recognition of cursive Nastalique Urdu script [17]. The hybrid approach extracts invariant features through Convolution Neural Networks (CNN) and then fed these features to MDLSTM layers for learning. Experiments are performed on UPTI dataset with an accuracy of 98.12%. This hybrid approach achieved a high recognition rate on Nastalique script.

V. CONCLUSION

Mainly, explicit character segmentation creates high challenges in Naskh and Nastalique text recognition process and affects the overall accuracy of the OCR system. Therefore, researchers in this field avoid the character segmentation phase and use segmentation free approaches for the recognition of Naskh and Nastalique text. In this article, a study has been made for the cursive Naskh/Nastalique text recognition by using deep neural network architectures, and it is concluded that RNN-LSTM based methods are best suitable for the cursive and context sensitive languages recognition as it avoids segmentation and handcrafted feature extraction, two hard phases of OCR process for cursive script. The LSTM based architecture can achieve accuracies beyond 90% for highly complex writing style such as Naskh and Nastalique based languages.

Table 1: Details of LSTM-RNN network for Naskh and Nastalique wrting style languages

Article	Language	Lines/words	Network type	Input Block	Hidden size	Feed-Forward size	Database	Character Error Rate
[7]	Pastho	Normalized Lines	BLSTM	1*48	120	--	Custom	16.16%
[7]	Pastho	Non-normalized Lines	MDLSTM	1*4	4,20,100	16,80	Custom	9.22%
[7]	Pastho	Normalized Lines	MDLSTM	1*4	4,20,100	16,80	Custom	9.33%
[12]	Arabic	Words	MDLSTM	3*4	2,10,50	6,20	APTI	12.09%
[13]	Urdu	Normalized Lines	MDLSTM	4*1	2,10,50	6,20	UPTI	2%

REFERENCES

[1] G.S. Lehal, "*Choice of recognizable units for urdu OCR*", In Proceedings of the Workshop on Document Analysis and Recognition (DAR'12), pp.**79–85**, **2012.**

[2] S. Garg, A.P. S, K. C, "*An Extensive Survey on Text Detection and Recognition*", International Journal of Computer Sciences and Engineering, Vol.**7**, No.1, pp.**546-551, 2019.**

[3] A. Ul-Hasan, S.B. Ahmed, F. Shafait, T.M. Breuel, "*Offine printed Urdu Nastaleeq script recognition with bidirectional LSTM networks*", In Proc. 12th Int. Conf. Document Analysis Recognition (ICDAR), pp.**1061-1065, 2013**.

[4] A. Ul-Hasan, M.Z. Afzal, F. Shafait, M. Liwicki, T.M. Breuel, "*A sequence learning approach for multiple script identification*", In Document Analysis and Recognition (ICDAR),pp.**1046-1050, 2015**.

[5] A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber, "*Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*", In Proceedings of the 23rd International Conference on Machine learning, pp.**369–376, 2006**.

[6] S.B. Ahmed, S. Naz, M.I. Razzak, S.F. Rashid, M.Z. Afzal, T.M. Breuel, "*Evaluation of cursive and non-cursive scripts using recurrent neural networks*", Neural Comput. Appl. vol. **27,** no. **3**, pp.**603-613, 2016.**

[7] R. Ahmad, Z. Afzal, S.F. Rashid, "*KPTI: Katib's Pashto Text Imagebase and Deep Learning Benchmark*", 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), **2016**.

[8] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, J. Schmidhuber, "*A Novel Connectionist System for Unconstrained Handwriting Recognition*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. **31,** No. **5**, pp.**855–868,2009.**

[9] S.F. Rashid, M. Schambach, J. Rottland, S. Nll, "*Low resolution arabic recognition with multidimensional recurrent neural networks*", In Proceedings of the 4th International Workshop on Multilingual OCR, **2013**.

[10] R. Ahmad, Z. Afzal, S.F. Rashid, M. Liwicki, T.M. Breuel, "*Scale and rotation invariant ocr for Pashto cursive script using mdlstm network*", In Document Analysis and Recognition (ICDAR), pp.**1101-1105**, **2015**.

[11] R. Ahmad, S. Naz, Z. Afzal, S.F. Rashid, M. Liwicki, A. Dengel, "*Deepkhatt: A deep learning benchmark on arabic script*", Document Analysis and Recognition (ICDAR). 14th International Conference on IEEE, **2017**.

[12] R. Maalej, N. Tagougui, and K. Kherallah, "*Recognition of Handwritten Arabic Words with Dropout Applied in MDLSTM*" ICIAR, pp.**746–752, 2016**.

[13] S. Naz, A. I. Umar, R. Ahmed, M.I. Razzak, S.F. Rashid, F. Shafait, "*Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks*", Springer-Plus, vol. **5**, no. **1**, pp.**1-16**, **2016**.

[14] S. Naz, A.I. Umar, R. Ahmad, S.B. Ahmed, S.H. Shirazi, I. Siddiqi, and M.I. Razzak, "*Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks*" Neurocomputing, vol. **177**, pp. **228-241, 2016**.

[15] S. Naz, A. I. Umar, R. Ahmad, S.B. Ahmed, S.H. Shirazi, M.I. Razzak, "*Urdu Nasta'liq text recognition system based on multidimensional recurrent neural network and statistical features*", Neural Comput. Appl. vol. **28**, no. **2**, pp. **219-231**, **2016**.

[16] S. Naz, S.B. Ahmed, R. Ahmad, M.I. Razzak, "*Zoning features and 2DLSTM for Urdu text-line recognition*", Procedia Computer Science Vol. **96**, No. **1**, pp.**16-22**, **2016**.

[17] S. Naz, A. I. Umar, R. Ahmad, I. Siddiqi, S.B. Ahmed, M.I. Razzak, F. Shafait, "*Urdu Nastaliq recognition using convolutional recursive deep learning*" Neurocomputing, vol. **243**, pp. **80-87**, **2017**.

Authors Profile

Shanky Goel received her Bachelors degree in mathematics and Post Graduate degree in Computer Science from Punjabi University, Patiala, India. She is pursuing Ph.D. from Punjabi University, Patiala, Punjab, India. Her research interests include Character Recognition and Natural Language Processing.

Professor Gurpreet Singh Lehal received undergraduate degree in Mathematics from Punjab University, Chandigarh, India, and Post Graduate degree in Computer Science from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, India. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and research. He is Director of Research Centre for Punjabi Language Technology and Dean of Faculty of Computing Sciences, Punjabi University, Patiala. His areas of research are- Natural Language Processing and Optical Character Recognition. He has published more than 100 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions- Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration.

