

Photometric Ligature Extraction Technique for Urdu Optical Character Recognition

Majida Kazmi

Faculty of Electrical and Computer Engineering
NED University of Engineering and Technology
Karachi, Pakistan
majidakazmi@neduet.edu.pk

Fauzia Yasir

Faculty of Electrical and Computer Engineering
NED University of Engineering and Technology
Karachi, Pakistan
fyasir@neduet.edu.pk

Samreen Habib

Neurocomputation Lab, NCAI
NED University of Engineering and Technology
Karachi, Pakistan
habib@cloud.neduet.edu.pk

Muhammad Saad Hayat

Department of Electrical Engineering
NED University of Engineering and Technology
Karachi, Pakistan
hayat@cloud.neduet.edu.pk

Saad Ahmed Qazi

Faculty of Electrical and Computer Engineering
Neurocomputation Lab, NCAI
NED University of Engineering and Technology
Karachi, Pakistan
saadqazi@neduet.edu.pk

Abstract-Urdu Optical Character Recognition (OCR) based on character level recognition (analytical approach) is less popular as compared to ligature level recognition (holistic approach) due to its added complexity, characters and strokes overlapping. This paper presents a holistic approach Urdu ligature extraction technique. The proposed Photometric Ligature Extraction (PLE) technique is independent of font size and column layout and is capable to handle non-overlapping and all inter and intra overlapping ligatures. It uses a customized photometric filter along with the application of X-shearing and padding with connected component analysis, to extract complete ligatures instead of extracting primary and secondary ligatures separately. A total of ~2,67,800 ligatures were extracted from scanned Urdu Nastaliq printed text images with an accuracy of 99.4%. Thus, the proposed framework outperforms the existing Urdu Nastaliq text extraction and segmentation algorithms. The proposed PLE framework can also be applied to other languages using the Nastaliq script style, languages such as Arabic, Persian, Pashto, and Sindhi.

Keywords-ligature; holistic; Urdu OCR; Nastaliq; photometric filter; Urdu printed text images

I. INTRODUCTION

OCR technology is used to obtain machine editable text from text images. It allows the digitization of valuable printed and handwritten data covering cultural and historical milestones [1]. The commercial OCR systems that are now available report near to 100% recognition rates for languages

using the Latin alphabet, such as English, German, and French. Arabic and Chinese OCR systems are also well-developed. Despite the significant research interest in this area, OCR systems for many languages, including Urdu, are still in the development stage [2-3]. Urdu is Pakistan's official language having a large collection of valuable printed and handwritten data in the form of books, novels, magazines, and newspapers. Most of these valuable data are not accessible digitally. The Urdu language has 39 basic characters, 28 of which are Arabic. It is mostly written in the Nastaliq script style, which is a complex calligraphic style, written diagonally from right-to-left with varying inter and intra word spaces, overlapping of characters and strokes, incorrect or filled loops and lack of fixed baseline [4-5] as shown in Figure 1.

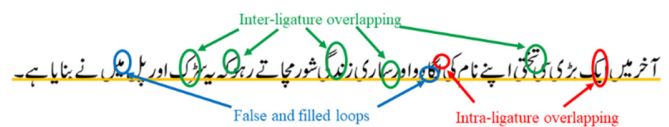


Fig. 1. Major challenges in Nastaliq text: Intra overlapping ligatures (red), inter overlapping ligatures (green), false and filled loops (blue) and missing baseline (yellow).

Urdu OCR is primary composed of five stages: Image acquisition, pre-processing, segmentation, classification and recognition, and post-processing [6]. Image acquisition collects digital images through camera shots, scanned text images, or

Corresponding author: Majida Kazmi

generated synthetic images [6]. Pre-processing aims to enhance the quality of an acquired image [6]. Noise and skew removal, binarization, contrast enhancement, etc. are mainly performed in this step with the use of classic image processing techniques. Segmentation decomposes a source image into characters, ligatures, or words [7-8]. This step usually employs projection profile and Connected Component Analysis (CCA). Classification aims to correctly classify the extracted/segmented features (ligatures, characters, words, etc.). The most common classifier methods are Decision Tree (DT), Statistical Classifier (SC), Neural Networks (NNs) [9, 10], Hidden Markov Models (HMMs), and Support Vector Machines (SVMs). Finally, post-processing corrects the recognition errors in the obtained text [10]. The techniques used for OCR post-processing include manual error correction, dictionary-based error correction, and context-based error correction [12-13].

Among the above stages, segmentation at character, ligature, or word level is the most challenging stage in Urdu OCR. Based on these levels, Urdu OCR can be divided into two categories: analytical approach at character level [14-15] and holistic approach at ligature level [7, 16-17]. The analytical approach segments text at character level either explicitly or implicitly. The explicit segmentation requires an extensive knowledge of characters as it explicitly divides handwritten or printed text into characters. Many researchers have adopted the explicit character segmentation [17-21]. On the other hand, implicit segmentation is an integration of the segmentation and recognition processes. Successful work has been reported by researchers for implicit segmentation [22-26] due to the smaller number of segments. However, both algorithms require a massive amount of training data for better results. The holistic approach is also referred to as segmentation-free method. It extracts at ligature or word level. Groups of isolated (non-joiner) characters and non-isolated (joiner) characters (Figure 2(a)) are termed as ligatures. These ligatures are grouped to form words. Ligatures are further classified as primary and secondary ligatures. Primary ligatures represent the main body of a word, while dots or diacritic marks are the secondary ligatures (Figure 2(b)).

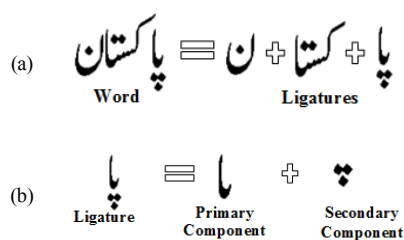


Fig. 2. Word breakdown. (a) Ligatures in an example word, (b) ligature components.

Avoiding character level segmentation has made the holistic method extremely popular [3, 27-32]. Authors in [27] followed the projection technique for text line extraction. The main body and diacritics were identified based on the distance between the horizontal base and the average line. The

technique was tested on a small data set that was not specific to Nastaliq script, consisting of 1050 single characters and ligatures, with 98.86% accuracy. Authors in [28] used the horizontal projection technique. CCA was applied before text segmentation. The horizontal span of each secondary component on the baseline was calculated for the re-association of diacritics to their respective primary ligature. However, this approach assumed to work on text files instead of text images to extract complete ligatures. Similarly, authors in [29] applied the vertical projection profile method for the association of secondary ligatures by calculating the start and end point of diacritics. The proposed method reported 100% and 99% accuracy in baseline identification and ligature extraction respectively on scanned images with 48 font size but this technique ignores intra-overlapping ligatures and is also font size dependent.

Authors in [3] employed the horizontal projection method along with dilation to merge secondary and primary ligatures before line separation from the image. Authors in [30] used only 300 ligature samples to evaluate their proposed method, reporting 91.3% accuracy in segmentation and 78% in diacritics association. Authors in [31] proposed an extraction ligature technique based on 6 heuristic conditions reporting an accuracy of 99.02% on 45 images. Authors in [32] proposed the line segmentation technique with the connected component analysis method on images to collect width, height and centroids of ligatures reporting 99.80% accuracy. However, this technique does not segment multi-column scripts and overlapped inter and intra ligatures. Many recognition techniques carry out separate classifications of primary and secondary components [3, 27-32] to reduce the number of distinct recognizable classes. Such techniques face significant challenges in re-associating the secondary components with their primary components to recognize the entire ligature. The complexity at character segmentation has shifted the focus towards the holistic approach, i.e. the recognition of words or ligatures in the text. Segmenting text at the character level is more complex than the recognition of words and ligatures due to character overlapping, varying inter and intra word spaces, context sensitivity, different forms of characters according to their position in a word or a ligature, and the cursive script style. The literature review reveals that Urdu OCR is an open field for the researcher to design a system capable of incorporating factors such as intra and inter ligature overlapping, multi-column text images with borders, font variation, and mass data of ligatures for classification.

An efficient ligature extraction technique for Urdu OCR is proposed in this paper. The proposed method is capable to extract complete ligatures efficiently unlike separating primary and secondary components. The proposed technique is independent of font size and column layout, and is capable to handle all overlapping and non-overlapping ligatures by addressing the issue of intra overlapped ligatures as well as the complex association of the secondary components. It extracts complete ligatures, rather than separating primary and secondary components, thus secondary ligatures do not need to reassociate with their primary ligature in the classification and recognition steps. The proposed framework is designed for

Urdu but is applicable to other languages that follow the Nastaliq style, such as Arabic, Persian, Pashto, and Sindhi.

II. THE PROPOSED METHODOLOGY

The proposed framework for ligature extraction is depicted in Figure 3. It consists of 3 steps: image acquisition, image binarization, and PLE. Urdu printed text images from novels, religious books written in Nastaliq style, in single and double columns and varying font sizes were downloaded from different sources [33] and are referred to as I_{img} . First, the I_{img} is converted into binarized images I_{th} by using hard thresholding. The resultant I_{th} is a mono-chrome image with white background and black text (Figure 4). Then, an efficient process of PLE is applied on each I_{th} .

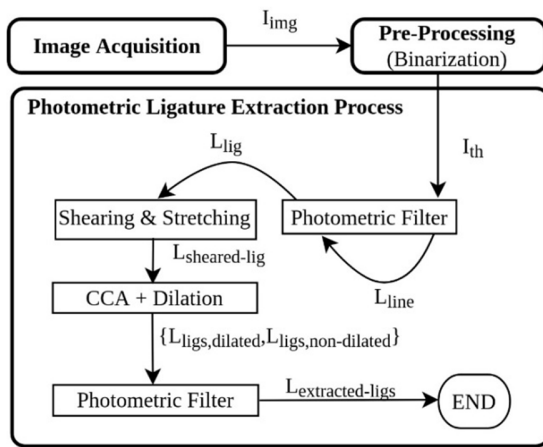


Fig. 3. Framework for Urdu ligature extraction.

A. Photometric Ligature Extraction (PLE)

The proposed PLE used a customized photometric filter which is specifically designed to decompose an image based on the photometric similarity. The stepwise description of PLE process follows:

- In the first step, PLE deploys a photometric filter to extract text lines (L_{lines}) from the image (I_{th}). The algorithm in Figure 4 demonstrates the working of the photometric filter. This filter scans the binarized image from top to bottom to detect text using the logical AND operator. The size of the photometric filter is adjusted with the width (W) of the image as $(1 \times W)$. The output of the photometric filter is then saved in an array. The resultant array is a stream of zeroes and ones, on which unary AND operation is performed to get a single bit value, i.e. 0 or 1. The 0 value indicates the presence of black pixel/s in the row, otherwise the value will be 1.
- In the second step, the image L_{lines} is first rotated counterclockwise by 90° . The photometric filter is then applied to each line of L_{lines} to extract both overlapped and non-overlapped ligatures.
- The overlapped ligatures are corrected in this step by applying X-shear transformation and padding simultaneously on each L_{lig} to overcome the most

challenging issue of inter and intra ligatures overlapping. The output of this step consists of the sheared and padded ligatures $L_{sheared-lig}$.

- In this step, the $L_{sheared-lig}$ images are classified into two classes based on the extent value of the first encountered ligature in image using CCA. Height, width, centroid, etc. are major properties obtained through the CCA method. The developed methodology utilized another component property termed as extent which is defined as the ratio of contour area to the bounding rectangle area. The extent value is a key feature in distinguishing secondary and primary ligatures with 99% accuracy. If the extent value of ligature is less than the hard threshold value, then dilatation operation is carried out on the encountered ligature producing $L_{ligs, dilated}$. This process reduces the distance between the primary and the secondary component of a ligature.
- In the last step, the photometric filter is again applied to all dilated and non-dilated ligatures $L_{ligs,dilated}$ and $L_{ligs,non-dilated}$ to extract complete ligatures as final output $L_{extracted-ligs}$.

```

Input: UrduPrintedTextImages { $I_{img}$ }
Output: SegmentedLines { $L_{lines}$ }
for each image do
    Rows=height of image, Set Index=0
    Apply thresholding & saved as binarized image
    Construct Photometric filter [ $1 \times W$ ] size
    Background pixel values are co-efficient of filter
    while Index < Rows do
        Apply AND operation on each row
        res=resAND.getAND(Filter, thresholdedImg,)
        if res==0 then
            startLine=getRowvalue(i)
            Index++
            while res!=0 do
                endLine=getRowvalue(i)
                Index++
            end
            cropLine=thresholdedImg[startLine:endLine]
            Append to the list of extracted line (Lline)
        end
        Index++
    Return extracted lines from image
    
```

Fig. 4. The photometric filter algorithm.

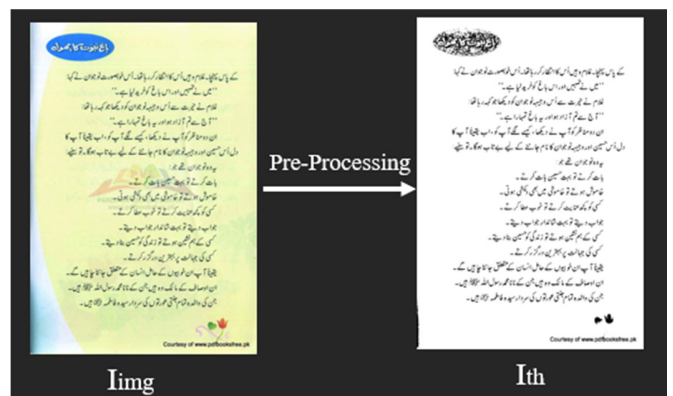


Fig. 5. The result of image binarization (I_{th}) on the input image I_{img} .

B. Demonstration of the Proposed PLE

The stepwise demonstration of the proposed PLE technique is shown in Figure 6. The input of the PLE technique is a mono-chrome image with white background and black text (Figure 5).

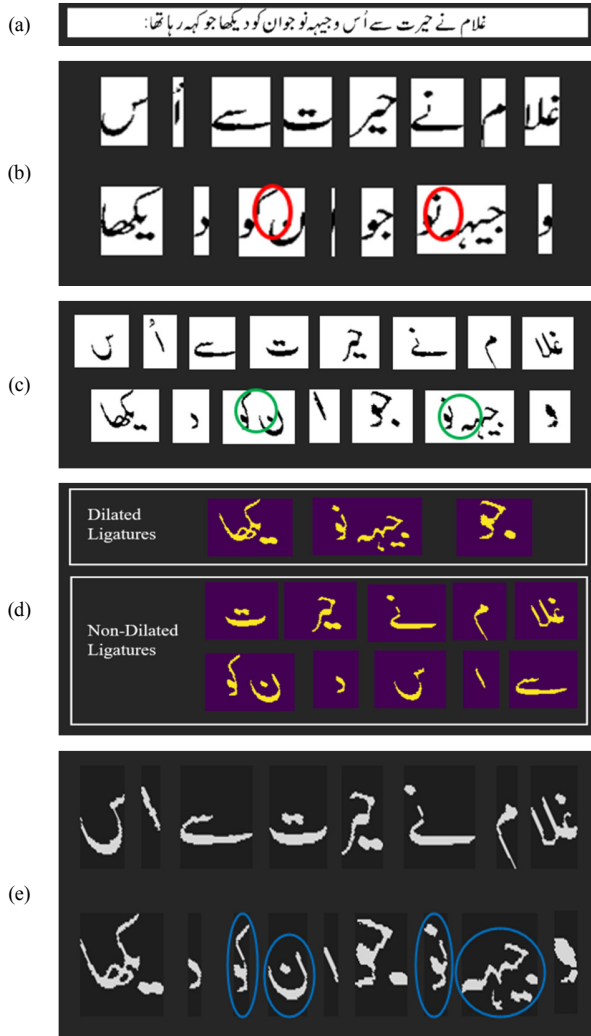


Fig. 6. PLE framework illustration. (a) An extracted line (L_{line}). (b) Segmented ligatures from the sentence in (a). The red encircled ligatures are overlapped. (c) The overlapping issue of ligatures obtained in (b) is resolved by X-shearing of ligatures encircled as green. (d) List of ligatures ($L_{liggs,dilated}, L_{liggs,non-dilated}$) after morphological operation (dilation). (e) Correctly extracted ligatures ($L_{extracted-ligs}$) after PF applied on ($L_{liggs,dilated}, L_{liggs,non-dilated}$).

In the first step, text lines are extracted one by one from the text image by applying the photometric filter (Figure 6(a)). In the next step, each extracted line is first rotated counterclockwise and then again passes through the photometric filter to extract both overlapped (marked as red circle) and non-overlapped ligatures (Figure 6(b)). The issue of inter and intra overlapping is resolved (see Figure 6(c), marked as green circles) by applying X-shearing and padding simultaneously on each ligature. Figure 6(d) depicts the list of

dilated and non-dilated ligatures. The dilation process reduces the distance between the primary and secondary component of a ligature. Finally, the photometric filter is again applied on these ligatures to get the final output as shown in Figure 6(e). This step will further enhance the correct separation of ligatures.

III. RESULTS AND ANALYSIS

The proposed Urdu ligature extraction framework was evaluated on downloaded Urdu printed text images. The technique was tested on a total of 600 novel and book images. The working dataset mainly comprised of non-overlapping lines with no boundary across images. First, the photometric filter was applied on the images and extracted lines with an accuracy of 99.6%. A total of 13,200 lines were extracted from 600 images. These lines were then segmented into ligatures. A total of 267,800 ligatures were extracted after the complete execution of all the steps of the proposed PLE with an overall accuracy of 99.4%. Table I compares the proposed ligature extraction framework with previously reported methods. Authors in [27] evaluated their approach on 1050 ligatures with 98.86% accuracy in primary and secondary stroke extraction. Authors in [29] achieved 99% accuracy in ligature and diacritics extraction. Authors in [30] tested their system on 300 sample images out of which 274 were segmented correctly with 91.3% accuracy. Authors in [30] analyzed 45 Urdu images to classify and associate the connected components with 99.02% accuracy. Authors in [32] used 10,063 text lines to test their algorithm and reported an accuracy of 99.8%.

However, as discussed above, due to the limited data set of ligatures, researchers have mostly deployed algorithms on their own datasets to check the accuracy of ligature segmentation/extraction. Therefore, the accuracy depends upon the complexity of the text images used for segmentation and re-association of primary and secondary components. Segmentation algorithms achieving segmentation accuracy near 99% apply CCA for primary and secondary component segmentation in [28, 31-32] and projection profile method in [3, 27, 29-30] and then reassociate the secondary components. These studies also ignore the extraction of inter overlapped ligatures. The last row of Table I presents the findings of the proposed technique. The proposed solution resolved the problem of inter ligature overlapping with an accuracy of 99.4%. However, the efficiency of the proposed method is reduced due to the redundant use of the word جو . This complete ligature remains unaffected even after X-shearing because the primary component Alif "ا" lies in the region of the second main component 'ک' and the diacritics also overlap with the neighboring primary components. It was observed that spacing between diacritics that lie below the main body sometimes leads to incorrect line segmentation.


IV. CONCLUSION

This paper presented an efficient ligature extraction technique for the extraction of Urdu ligatures in Nastaliq fonts. The technique used a customized photometric filter along with the application of X-shearing and padding with CCA that result in the efficient extraction of overlapped and non-overlapped ligatures. The proposed framework achieves an accuracy of

99.4%. The efficiency of PLE technique can be enhanced by overcoming the association of secondary ligatures to respective main component before the extraction of text lines. This work

can also be deployed for other Nastaliq script-based languages like Persian, Pashto, Saraiki, Panjabi, etc..

TABLE I. COMPARISON WITH RELEVANT HOLISTIC APPROACHES

Work	Ligature Extraction		Limitations	Accuracy (%)	Extracted overlapped ligatures
	Technique	Result			
[32]	CCA	Separate primary and secondary components	Declares the secondary ligature as primary ligature when the size exceeds the threshold value	99.8	No
[30]	PP	Separate primary and secondary components	- Extracted only 300 ligatures - Low diacritics association accuracy	91.3	No
[31]	CCA	Separate primary and secondary components	- Relies on zonal information - Tested on only 45 images	99.02	No
[29]	PP	Separate primary and secondary components	- Cannot extract overlapped ligatures - Dependent on font size of 48	99.0	No
[27]	PP	Separate primary and secondary components	- Did not specify the script style - The system was tested on 1050 ligatures	98.86	No
[28]	CCA	Separate primary and secondary components	The proposed method was directly tested on text files	97.4	No
Proposed PLE	Photometric filter, X-shearing and stretching, CCA, and dilation	Complete ligature extraction	- Uneven baseline recognition - Narrow spacing between the ligatures of the word  which reduces efficiency	99.4	Yes

CCA: Connected Component Analysis, PP: Projection Profile

REFERENCES

- [1] A. Wali and S. Hussain, "Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation," 2007, pp. 53–58, https://doi.org/10.1007/978-1-4020-6268-1_10.
- [2] S. T. Javed and S. Hussain, "Segmentation Based Urdu Nastaliq OCR," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 2013, pp. 41–49, https://doi.org/10.1007/978-3-642-41827-3_6.
- [3] I. U. Din, Z. Malik, I. Siddiqi, and S. Khalid, "Line and Ligature Segmentation in Printed Urdu Document Images," presented at the 3rd International Conference on Computational and Social Sciences, Oct. 2015.
- [4] S. Naz, A. I. Umar, S. B. Ahmed, S. H. Shirazi, M. Imran Razzak, and I. Siddiqi, "An Ocr system for printed Nasta'liq script: A segmentation based approach," in *17th IEEE International Multi Topic Conference 2014*, Dec. 2014, pp. 255–259, <https://doi.org/10.1109/INMIC.2014.7097347>.
- [5] H. R. Khan, M. A. Hasan, M. Kazmi, N. Fayyaz, H. Khalid, and S. A. Qazi, "A Holistic Approach to Urdu Language Word Recognition using Deep Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7140–7145, Jun. 2021, <https://doi.org/10.48084/etasr.4143>.
- [6] N. H. Khan and A. Adnan, "Urdu Optical Character Recognition Systems: Present Contributions and Future Directions," *IEEE Access*, vol. 6, pp. 46019–46046, 2018, <https://doi.org/10.1109/ACCESS.2018.2865532>.
- [7] S. Chanda and U. Pal, "English, Devnagari and Urdu Text Identification," in *Proc. international conference on document analysis and recognition*, 2005, pp. 538–545.
- [8] A. Rana and G. S. Lehal, "Offline Urdu OCR using Ligature based Segmentation for Nastaliq Script," *Indian Journal of Science and Technology*, vol. 8, no. 35, pp. 1–9, Dec. 2015, <https://doi.org/10.17485/ijst/2015/v8i35/86807>.
- [9] M. Alghobiri, "A Comparative Analysis of Classification Algorithms on Diverse Datasets," *Engineering, Technology & Applied Science Research*, vol. 8, no. 2, pp. 2790–2795, Apr. 2018, <https://doi.org/10.48084/etasr.1952>.
- [10] S. R. Basha, J. K. Rani, and J. J. C. P. Yadav, "A Novel Summarization-based Approach for Feature Reduction Enhancing Text Classification Accuracy," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 5001–5005, Dec. 2019, <https://doi.org/10.48084/etasr.3173>.
- [11] I. A. Doush, F. Alkhateeb, and A. H. Gharaibeh, "A novel Arabic OCR post-processing using rule-based and word context techniques," *International Journal on Document Analysis and Recognition (IJ DAR)*, vol. 21, no. 1, pp. 77–89, Jun. 2018, <https://doi.org/10.1007/s10032-018-0297-y>.
- [12] Y. Bassil and M. Alwani, "OCR Post-Processing Error Correction Algorithm using Google Online Spelling Suggestion," *arXiv:1204.0191 [cs]*, Apr. 2012, Accessed: Dec. 01, 2021. [Online]. Available: <http://arxiv.org/abs/1204.0191>.
- [13] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys*, vol. 24, no. 4, pp. 377–439, Dec. 1992, <https://doi.org/10.1145/146370.146380>.
- [14] S. Naz, K. Hayat, M. Imran Razzak, M. Waqas Anwar, S. A. Madani, and S. U. Khan, "The optical character recognition of Urdu-like cursive scripts," *Pattern Recognition*, vol. 47, no. 3, pp. 1229–1248, Mar. 2014, <https://doi.org/10.1016/j.patcog.2013.09.037>.
- [15] S. A. Husain, "A multi-tier holistic approach for Urdu Nastaliq recognition," in *International Multi Topic Conference, 2002. Abstracts. INMIC 2002.*, Karachi, Pakistan, Dec. 2002, pp. 84–84, <https://doi.org/10.1109/INMIC.2002.1310191>.
- [16] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil, and H. Moin, "Segmentation Free Nastaliq Urdu OCR," *International Journal of Computer and Information Engineering*, vol. 4, no. 10, pp. 1514–1519, Oct. 2010.
- [17] U. Pal and A. Sarkar, "Recognition of printed Urdu script," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, Edinburgh, UK, Aug. 2003, pp. 1183–1187, <https://doi.org/10.1109/ICDAR.2003.1227844>.
- [18] Z. Ahmad, J. K. Orakzai, and I. Shamsheer, "Urdu compound Character Recognition using feed forward neural networks," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, Beijing, China, Aug. 2009, pp. 457–462, <https://doi.org/10.1109/ICCSIT.2009.5234683>.
- [19] S. A. Sattar, S. Haque, and M. K. Pathan, "A Finite State Model for Urdu Nastaliq Optical Character Recognition," *International Journal of Computer Science and Network Security*, vol. 9, no. 9, pp. 116–122, 2009.
- [20] S. T. Javed, "Investigation into a segmentation-based OCR for the Nastaleeq writing system," M.S. thesis, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2007.
- [21] S. Mir, S. Zaman, and M. W. Anwar, "Printed Urdu Nastaliq Script Recognition Using Analytical Approach," in *2015 13th International*

- Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, Dec. 2015, pp. 334–340, <https://doi.org/10.1109/FIT.2015.65>.
- [22] S. B. Ahmed, S. Naz, M. I. Razzak, S. F. Rashid, M. Z. Afzal, and T. M. Breuel, "Evaluation of cursive and non-cursive scripts using recurrent neural networks," *Neural Computing and Applications*, vol. 27, no. 3, pp. 603–613, Apr. 2016, <https://doi.org/10.1007/s00521-015-1881-4>.
- [23] R. P. Thakkar Mitesh, "Handwritten Nastaleeq Script Recognition with BLSTM-CTC and ANFIS method," *International Journal of Computer Trends and Technology*, vol. 11, no. 3, 2014, <https://doi.org/10.14445/22312803/IJCTT-V11P128>.
- [24] S. Naz *et al.*, "Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks," *Neurocomputing*, vol. 177, pp. 228–241, Feb. 2016, <https://doi.org/10.1016/j.neucom.2015.11.030>.
- [25] S. Naz *et al.*, "Urdu Nastaliq recognition using convolutional–recursive deep learning," *Neurocomputing*, vol. 243, pp. 80–87, Jun. 2017, <https://doi.org/10.1016/j.neucom.2017.02.081>.
- [26] S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi, and M. I. Razzak, "Urdu Nastaliq text recognition system based on multi-dimensional recurrent neural network and statistical features," *Neural Computing and Applications*, vol. 28, no. 2, pp. 219–231, Feb. 2017, <https://doi.org/10.1007/s00521-015-2051-4>.
- [27] S. Sardar and A. Wahab, "Optical character recognition system for Urdu," in *2010 International Conference on Information and Emerging Technologies*, Karachi, Pakistan, Jun. 2010, <https://doi.org/10.1109/ICIET.2010.5625694>.
- [28] N. Sabbour and F. Shafait, "A segmentation-free approach to Arabic and Urdu OCR," *Proc. SPIE*, vol. 8658, p. 86580N, Feb. 2013.
- [29] S. Nazir and A. Javed, "Diacritics Recognition Based Urdu Nastalique OCR System," *The Nucleus*, vol. 51, no. 3, pp. 361–367, Sep. 2014.
- [30] A. F. Ganai and A. Koul, "Projection profile based ligature segmentation of Nastaleeq Urdu OCR," in *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*, Olten, Switzerland, Sep. 2016, pp. 170–175, <https://doi.org/10.1109/ISCBI.2016.7743278>.
- [31] G. S. Lehal, "Ligature Segmentation for Urdu OCR," in *2013 12th International Conference on Document Analysis and Recognition*, Washington, DC, USA, Aug. 2013, pp. 1130–1134, <https://doi.org/10.1109/ICDAR.2013.229>.
- [32] I. Ahmad, X. Wang, R. Li, M. Ahmed, and R. Ullah, "Line and Ligature Segmentation of Urdu Nastaleeq Text," *IEEE Access*, vol. 5, pp. 10924–10940, 2017, <https://doi.org/10.1109/ACCESS.2017.2703155>.
- [33] "Kutubistan," *Kutubistan*. <https://kutubistan.blogspot.com/> (accessed Dec. 01, 2021).