International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. comp. info. Vol. 3 (2024) 32–43

Paper Type: Original Article

Urdu Handwriting Recognition with Deep Learning: Current Methods and Future Prospects

Safaa Saber [1] **and Mohamed G. Mahdi** [1,2,*]

[1] Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt. Emails: s.saber22@fci.zu.edu.eg; m.gresha@fci.zu.edu.eg.

[2] Department of Computer Science, Higher Institute for Computer Sciences and Information Systems, 5th Settlement, New Cairo, Egypt; mohamed.grisha@cis.edu.eg.

Received: 15 Jan 2024 **Revised**: 28 Apr 2024 **Accepted**: 12 May 2024 **Published**: 15 May 2024

Abstract

The Nastaliq script's calligraphic and cursive features—where letter forms change based on where they appear in a word—make Urdu handwriting recognition a challenging process. Urdu language has not been studied in this way except for a few languages such as English, Arabic, where some trends in handwriting recognition have been noted. This review is concerned with trends in deep learning abstraction, especially convolutional neural networks (CNNs) which have shown success in handwritten text recognition of Urdu. It has been highlighted though that there is a number of significant challenges such as the complex nature of the ligatures, diversity of the writing tendencies, as well as insufficient quantity of extensive annotated Urdi datasets. In spite of the above context, some research in the address recognition systems has taken advantage of an Urdu-Nastaleeq Handwritten Dataset (UNHD) and Urdu Handwritten Text Dataset (UHTD), as well as a Urdu Handwritten Character Database (UHCD) but the research works still remains unreliable since the datasets are limited. Such approaches enable the researcher to make some drawing similarities not only between Urdu language and Pashto language but also with printed Persian Language. It is evident in this review that the center of attention has been the recognition of Urdu handwritten text rather than all the main points including language and algorithms.

Keywords: Urdu Handwriting Recognition; Deep Learning; Offline Handwriting Recognition; Writer Identification.

1 | Introduction

The handwritten text is an individual trait that differs from one person to another, thus making it indispensable in many identification and recognition applications [1–3]. Applications in writer identification and handwriting recognition find demand in several areas, including but not limited to, verification of historical documents, signature verification, protection of network access, and documents of a legal nature [4–9]. Two forms of handwriting can be distinguished: online handwriting, which refers to inserting text via different devices such as tablets and mobile phones, and offline handwriting, where a person writes on paper and later converts the writing to a computer in some form [8, 10, 11]. Regarding offline handwriting recognition, the writer notes that it is difficult for some complicated manipulations that concern scribes like nodded Urdu, to be carried out [11].

Corresponding Author: m.gresha@fci.zu.edu.eg

Over the recent past, deep learning and convolutional networks (CNNs) have made strides in handwriting recognition with impressive results, especially for the English and Arabic languages [12–15]. These models perform feature extraction from handwritten text automatically, which makes it unnecessary to employ hand-engineered features, one of the drawbacks of earlier recognition systems. Despite the combined similarities in the scripts of the two languages, the attention has been more so on Arabic than the development of deep learning-based systems for Urdu handwriting recognition [16, 17].

Offline Urdu handwriting recognition is still among the challenging issues [18, 19] due to the following reasons:

- The Nastaliq style is very calligraphic. Most of the time, letters are combined into complex ligatures, which sometimes makes the segmentation of individual characters nearly impossible for recognition.
- Different people have different ways of forming letters as part of their handwriting in Urdu. All those variations in stylistic flourishes and letter formation become a great bottleneck for its correct recognition.
- There are large datasets available for handwritten documents in English and Arabic, but the same cannot be said for Urdu. Hence, the offline recognition tasks, particularly for Urdu, have few comprehensive databases.
- In Urdu, each character changes its form depending on where it occurs within a word; this makes recognition much more difficult than in non-cursive scripts.

In this paper, we conduct a survey on the state-of-the-art in Urdu handwriting recognition with an extra focus on deep learning-based solutions. This paper has taken a brief overview of different methods that have been used to recognize Urdu handwriting, along with the classical machine learning and deep neural network-based approaches. The paper elaborates on how the development by the research community did not proceed well to develop a high-performance Urdu recognition system due to the required standard and large size of data as well as the adaptation of success models in the retrieval of other languages. This survey provides an extensive review of the literature and identifies inadequacies in state-of-the-art techniques to inspire further work centered on Urdu handwriting recognition while also helping raise awareness about this important research domain.

The remainder of the paper is organized as follows: Section 2 provides an overview of prior research that used deep learning to recognize the Urdu language. Section 3 discusses previous research on deep learning approaches applied to the identification of languages other than Urdu, such as Arabic. Finally, conclusions based on the preceding are offered in Section 4.

2 | Urdu Natural Language Processing

Urdu is one of the Indo-Aryan languages, and its cultural and social importance touches heights that only history can cover. It is the official language of Pakistan and one of the 22 scheduled languages of India, as an estimated 231 million people speak this language across the world. Estimates in 2022 by Ethnologue place Urdu behind the tenth-most spoken language in the world. It is very closely related to Hindi and has been sharing a significant portion of its grammar and vocabulary but with completely different scripts. Hindi is written in Devanagari, and, conversely, Urdu uses the Nastaliq script, which can be considered something similar to Arabic.

The right-to-left, cursive script of Urdu is especially distinctive. Urdu is written in the Nastaliq script, a form of Perso-Arabic-inspired calligraphy that is stylish and flows seamlessly from one letter to the next. The physical shape of written Urdu can be extremely different from its contextually dependent position in a word, much more so than many other languages. Every letter gets up to four shapes (initial, middle, end, standalone). This property, coupled with the existence of ligatures (where letters are fused), renders Urdu as an extremely complex script for Optical Character Recognition (OCR) and handwriting recognition systems.

2.1 | The Structure and Complexity of the Urdu Script

There are 38 letters and 10 numerals in the Urdu alphabet. Of these 38 characters, the 27 letters can be joined, as shown in Figure 1. The non-joiners create further difficulties by introducing extra segmentation and recognition processes. Since the script includes diacritical marks as well, which alter the pronunciation and thereby the meaning of the words, they also need to be identified with precision to achieve handwriting recognition successfully.

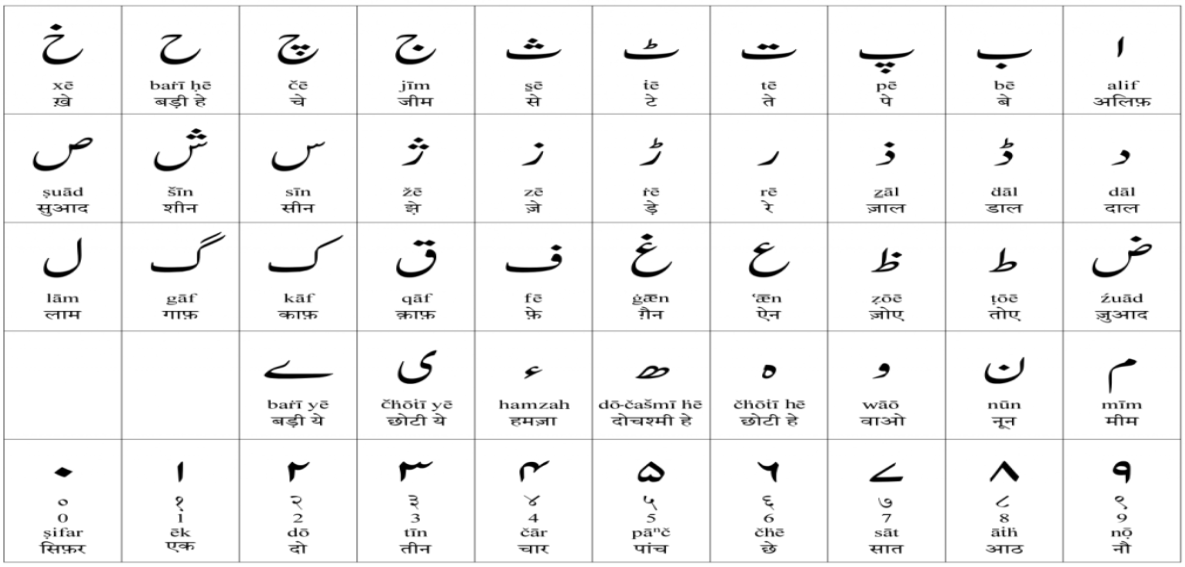

Figure 1. Alphabets of Urdu.

The Urdu script presents a challenge due to its varying letter shapes based on their position in a word (midway, end, or alone). The complexity of the morphological changeability for each character makes it harder for handwriting recognition systems to determine the changeability correctly. The already connected characters, making ligatures, obscure any sharp boundary between individual letters written in Urdu, adding to the degree of challenge to the segmentation and recognition processes. Distinctive features of the Urdu script make the character identification techniques consider variations both in forms and complexities of ligatures.

2.2 | Cultural and Linguistic Importance of Urdu

Urdu is a language of culture and literature, a vehicle for expressions of thought. It has a long history of great poets who brought the literature of the language to life through their philosophy, prose, and poetry, some of them such as Mirza Ghalib and Allama Iqbal. This is further associated with the cultural expression related to the elegance of the Nastaliq script, the so-called "calligraphy of the East." Hence, Urdu stands out somewhat as a very important language of cultural and intellectual influence in South Asia and will be a necessary tool for language tec and digital preservation.

The language of Urdu merits critical attention not only for its status as the official language of Pakistan but also for its predominant use among the Muslim population in India. Urdu is, with the unquestionable status of a literary and learned language, an apt twister for the disciplines of NLP and handwriting recognition on digitization and access to Urdu manuscripts and documents, contributing significantly toward their preservation and access.

2.3 | Challenges of Urdu Handwriting Recognition

The complexities of the Urdu script make the task even more challenging for handwriting recognition systems. Significant improvements in OCR and handwriting recognition have been made for languages such as English and Arabic, but the uniqueness of Urdu, especially the Nastaliq writing style, poses many challenges to developing efficient recognition systems. Writing from right to left and requiring the recognition of complex

ligatures, segmentation algorithms are often unsuccessful in discriminating between individual characters in texts or the influence of scripting on handwritten text [20, 21].

Another grave challenge is the unavailability of any sufficiently large standardized snapshot databases in Urdu handwriting recognition. The presence of large datasets in training deep learning models has proven extremely useful since deep learning models outperform traditional machine learning in pattern recognition. However, the majority of datasets are still focused on English, Arabic, and Chinese, with relatively few resources devoted to Urdu. Due to the underlying scarcity of data, such models that could have achieved very high accuracy for Urdu handwriting recognition have very little hope of being developed, especially considering the expectation of extremely high levels of accuracy with the newest kinds of handwriting recognition systems functioning offline-dependent upon the capturing of handwritten text and conversion to digital or electronic formats [22, 23].

2.4 | Future Directions and Importance in Technology

Big demands are posed by this task because governments, educational institutes, or firms are digitizing their records; hence, there is a need for 'hands-free' system processes, which would involve accurate document processing and recognition. This may vary from archiving historical manuscripts to automating forms in government offices or banks. While deep learning, especially convolutional neural networks (CNNs), may salvage some aspects of these difficulties that Urdu cursive handwriting poses, it must be said that CNNs may have been able to present themselves differently. CNNs proved immensely successful in dealing with a significant degree of variability and complexity of cursive scripts like Arabic and Persian, perniciously similar to Urdu structurally. The information that CNNs managed to extract automatically underpins their capability to extract subtle patterns in handwritten text to better augment recognition performances [24, 25]. To redress the serious shortcomings of less perfectly usable Urdu datasets, synthetic data reference generations and the application of transfer learning from corresponding languages like Arabic could find an effective role while training the models on a small but real-world Urdu dataset.

2.5 | Datasets for Urdu Handwriting Recognition

A large and diverse crowd-sourced dataset is one of the key elements for building efficient recognizers. Developing machine learning models for offline handwriting recognition has absolutely no chance unless we have a large enough dataset to train them. On one hand, Urdu's Nastaliq script presents an even greater challenge than Swir to its high level of cursiveness and calligraphy, but, on the other hand, the crowd-sourced datasets for learning Nastaliq script remain a bottleneck for Urdu handwriting recognition. There are some Urdu datasets developed over the past few years, although compared with the ones available in other languages, such as English or Arabic, their development is not as much.

2.5.1 Urdu-Nastaleeq Handwritten Dataset (UNHD)

The Urdu-Nastaleeq Handwritten Dataset (UNHD)[1] is the primary data resource for offline Urdu handwriting recognition. This dataset consists of thousands of samples of isolated characters, words, and sentences collected from various Urdu writers and covering gender and age groups. Secondly, the dataset covers the major constraints in the recognition of Nastaliq script, including links between letters and different forms of characters during different phases of writing. The dataset has been utilized in the training of deep neural networks—particularly convolutional neural networks (CNNs)—for automatic feature extraction and recognition tasks. However, UNHD is limited in size, and the dataset development is reflective of the urgency to create larger Urdu datasets to improve the recognition accuracy of Urdu handwriting. However, it can't be

[1] https://www.kaggle.com/datasets/drsaadbinahmed/unhd-dataset

downloaded directly from any official link. The dataset is only cited in academic publications. One might need to write an email to one or all of the authors of the dataset to get access to the dataset[26].

2.5.2 Urdu Handwritten Text Dataset (UHTD)

Another contribution in this regard is the UHTD[1], which will form the core assistance in improving the performance of Offline Urdu handwriting recognition systems. It is a set of many handwritten samples-from isolated characters to words and sentences-contributed by various writers for diversity. UHTD is structured to reflect variations in natural handwrites concerning speed, slant, and stroke thickness, which is very important for training deep learning models such as CNNs. The dataset has been utilized for word segmentation, character classification, and full-text recognition on handwriting recognition. While this database is a leap forward, it is not as large compared with datasets compiled for languages that enjoy much better-developed systems of recognition. To date, this UHTD dataset is not publicly available utilizing some specific online link. Access is usually provided by academic channels[28].

2.5.3 Urdu Handwritten Character Database (UHCD)

UHCD[2] is a dataset designed for capturing and representing individual characters and digits written in Urdu. The handwritten samples included run into thousands, collected from a wide demographic to ensure that varied writing styles are included. UHCD is very useful in character recognition tasks, where models have to classify individual Urdu characters or digits. It has gained a wide application in both traditional machine learning and deep learning models, where the system needs first to recognize isolated characters before dealing with full words or sentences. This dataset is very suitable for training CNNs and other neural networks concerning character-level handwritten recognition because of its well-annotated samples. The UHCD dataset is usually used for academic purposes, and access is provided after correspondence with the creators or relevant institutions. There is no public link directly to this dataset.[33, 34].

2.6 Related Work

In [35], a model based on deep neural networks is presented for the handwritten Urdu character recognition. A very large number of the system, 74,285, samples of input were placed into training and 21,223, samples of output, into trials. They achieved such a high recognition rate of 98.82% with 133 classes that the model exceeded the best current state-of-the-art systems. Also, it won an average recognition accuracy of 99.26% at a number of the numeral datasets of five languages. Besides, it had a 99.29% precision at the level of each language and a 99.322% accuracy at the general level.

In [36], the suggested model is built around two hotshot deep learning techniques, the CNN (Convolutional Neural Network) and the MDLSTM (Multi-dimensional Long Short-Term Memory) networks and it is solved for the recognition of the cursive Urdu Nastaliq script. The CNN locates low-level features that are used by the MDLSTM to pick up a deep understanding of the situation. On the UPTI dataset, the new approach outshone the prior assignments and, consequently, raised the standard to 98.12% correctness over 44 classes.

In [37], a hybrid deep learning model which consists of an encoder-decoder architecture is suggested. This system uses a CNN for feature extraction, with a bi-directional Gated Recurrent Unit (BiGRU) as the encoder and a Gated Recurrent Unit (GRU) as the decoder to recognize Urdu printed in the Nastaleeq font. With the dataset split into 50% for training, 30% for validation, and 20% for testing, the model achieved 98.5% accuracy on the test set, considering both 191 unique character-position categories and 99 basic categories.

In [38] Conv-transformer architecture. It is offered for unlimited offline Urdu handwriting recognition. The model combines CNN with a vanilla transformer, where CNN reduces spatial resolution to address challenges

[1] https://www.kaggle.com/datasets/saurabhshahane/urdu-handwritten-text-dataset
[2] https://www.kaggle.com/datasets/surindersinghkhurana/handwritten-urdu-characters-dataset

with the transformer's multi-head attention layer. They are trained on printed and handwritten lines of Urdu text. The model achieved a color error rate (CER) of 5%, but the authors note that additional CNN-based training data is needed to fully explore the potential of the classless transformer architecture.

In [39], an implicit segmentation method using the MDLSTM network was introduced to recognize Urdu text Nastalik. This system outperforms existing Urdu text line recognition models. With a recognition accuracy of 98±0.25%.

In [40], researchers presented the Urdu text line recognition system Nastalik using implicit segmentation and MDLSTM RNNs with an output layer of conjoint temporal classification (CTC). This method uses a window. Overlapping scrolls for feature extraction This leads to a 94.5% recognition rate in the UPTI database.

In [41], a system using multidimensional recurrent neural networks and statistical features was introduced to recognize Urdu text Nastalik through a three-step process: pre-testing and feature extraction, MDLSTM processing, and execution. The CTC was tested on the UPTI dataset and achieved 91.5% accuracy. The authors identified challenges such as the inability to handle scale variation. Limited diversity of datasets and the lack of comparative analysis with other state-of-the-art systems.

3 |Other Languages

Alongside that, Farsi, Kurdish, Arabic, and Pashto have obtained words, features, and structural elements from Urdu.

3.1 Arabic Language

Arabic is the world's fifth most widely spoken language, with over 420,000,000 people speaking it throughout the Middle East and North Africa. This is a Semitic language that uses the western branch of the South Semitic script. It is also written in Arabic script as is. The Arabic script has also been employed in several other languages such as Persian, Urdu, and Pashto [42]. Arabic script is cursive and written from right to left; most Arabic letters occur in four different forms depending on whether they are attached to an initial, medial, or final structure of a word, or whether the letter stands alone [43, 44].

Arabic handwriting recognition online and offline presents several significant challenges because of the many unique characteristics of the language. First, the cursive nature of Arabic writing makes segmentation into individual characters difficult. Many letters in Arabic are joined in ligatures, and the same letter may appear very differently depending on its position in a word. Besides, the diacritical marks showing the vowel sound may well change completely the meaning of words; therefore, their proper detection and recognition become highly relevant.

Deep learning techniques, especially using CNNs, have shown quite an important enhancement of performance in Arabic handwriting recognition in recent years. This ability of CNNs to learn the features from the data automatically turns them into effective handlers of the script's complexity. Most methods of machine learning require manually extracted features; a CNN can learn the detailed variations in handwritten Arabic text independently. Such models have achieved very high accuracy rates on both printed and handwritten Arabic text recognition [45].

However, some challenges have been faced while developing a robust Arabic recognition system:

- Arabic letters are variable in shape, depending on their position in the word. This adds more difficulty to the character recognition process that systems perform.
- Although some Arabic handwriting recognition datasets are available, they are mostly scarce in number and limited in their scope, especially for offline recognition.

- The need for their correct identification and interpretation adds to the complexity of the recognition process due to a high rate of recognition error from missing or wrong interpretation of a diacritical mark.

The authors of [46] propose a two-step approach for detecting and recovering out-of-vocabulary words in Arabic handwritten text recognition. They compare its effectiveness with one-step methods that rely either on a large static lexicon or sub-word modeling techniques. The results showed that their two-step approach performs the best; they obtained an accuracy of 91.5% for the detection of out-of-vocabulary words and 87.3% for recovery. This research thus gives very valuable insight into improving systems of Arabic handwriting recognition by overcoming out-of-vocabulary word challenges.

The authors in [47] present an Arabic script recognition system based on deep learning. The presented system is tested on the KHATT dataset. The proposed system is based on an MDLSTM architecture with a CTC layer for alignment. An augmentation method is used to enhance the input features. Results of an accuracy of 80% are reported on the KHATT dataset, which notably outperforms the earlier methods. The improvement is attributed to using deep learning along with data augmentation.

The authors in [48] presented a new approach incorporating the LSTM network with EHO. It uses hybrid features descriptors on the segmented character such as ELBP and IDMN. Feature dimensions are optimized by the EHO algorithm for reduction of the dimension of features over the improvement of over-fitting which enhances the training and testing of the classifier. These optimized features are then fed into the LSTM network for character classification. Simulation results are presented which show that the EHO-LSTM model yields an accuracy rate of 96.66%, 96.67%, and 99.93% for recognition of English, Kannada, and Arabic characters, respectively, on the Chars74K and MADbase digits dataset.

The authors in [49] propose a new model for recognizing both single-font and multi-font Arabic text by using the main classifiers of Support Vector Machine (SVM) and Convolutional Neural Network (CNN). In the proposed model, overfitting has been avoided embedding techniques of dropout and it automatically classifies and extracts features. A new training rule for deep neural networks is proposed using max-margin minimum classification error (M3CE) together with cross-entropy methods to enhance performance. The investigated model has been tested on more than one database and compared to most state-of-the-art methods in Arabic text recognition. Consequently, it turned out to be very effective and favorable in character recognition tasks.

Authors in [50] introduce a new method in the classification of handwritten Arabic characters using a Convolutional Neural Network combined with an optimized leaky ReLU activation function. These results are indicative that this approach is better in terms of improving accuracy, outperforming the standard ReLU at 97.8% and leaky ReLU at 97.9%, for all four datasets tested. This paper emphasizes that deep learning techniques greatly enhance performance related to the recognition of handwritten Arabic characters compared to the traditional approach and other activation functions such as ReLU and leaky ReLU.

Elkhayati and Elkettani [51] introduce a directed CNN model, which they call UnCNN. The proposed model is targeted for the recognition of isolated Arabic handwritten characters. Regarding this, the efficiency of the designed model was tested against BsCNN and other state-of-the-art techniques. Compared with many benchmark databases, such as IFHCDB, AHCD, AIA9K, and HACDB, the results have shown competitive performances using the UnCNN model, outperforming a lot of recent models in the literature, and hence proving its potential in effectively recognizing Arabic handwritten characters.

Yet despite these challenges, this development does not bring things to a standstill, and Arabic continues to be one of the languages for which much development in handwriting recognition systems is being achieved. Success with Arabic has also influenced the recent development of recognition systems in languages with the same script, such as Urdu and Pashto.

3.2 Pashto Language

Pashto, like Urdu and Arabic, uses a variant of the Arabic script, but it belongs to the Eastern Iranian branch of the Indo-Iranian languages. It is one of the two official languages of Afghanistan and is also widely spoken in parts of Pakistan, with over 40 million speakers globally. Pashto's script shares many similarities with Arabic and Persian scripts but also contains additional characters specific to the language [52–54].

The extra letters created by the cursive nature of Pashto writing add to the complication level and make recognition difficult for any handwriting system. Similar to other Arabic script languages, some issues with character segmentation regarding ligatures and different forms of the letter depending on their position in a word are attached to Pashto handwriting recognition. These are challenges that have made it very difficult for systems to identify the script correctly and interpret it, especially in handwritten texts where individual writing styles vary greatly.

3.2.1 Key Challenges in Pashto Handwriting Recognition

Extended Alphabet: Pashto has taken the Arabic script as a basis, but it includes other characters in the language. These are represented both by diacritics and unique consonants. Due to such aspects, Pashto handwriting recognition becomes more challenging, and recognition systems provided for Arabic or Persian do not perform well. Thus, separate models have to be designed specifically for Pashto.

Cursive Nature and Ligatures: The Pashto script is highly cursive, just like Arabic, meaning that in most cases, many letters within words are connected. This makes character segmentation quite problematic for handwriting recognition systems. For effective recognition, there should be correct identification of where a character ends and another begins, which may not be so easy considering the fluidity of Pashto handwriting.

Character Variability: Like Arabic and Urdu, the shape Pashto letters take depends on where in the word they are placed - Initial State, medial state, final state, or an isolated state. Not only this, but the difference in the shapes in these positions adds more to the misery because one letter can take many forms, which a handwriting recognition system must recognize for the proper classification of characters.

Lack of Large Datasets: Similar to Urdu, large and publicly available datasets relating to Pashto Handwritten Recognition, especially for offline recognition, barely exist. This absence of data hugely hampers the use and training of deep learning models since these models require large datasets to train well and achieve high accuracy. Filling the gap by creating synthetic data and transferring learning from related languages such as Arabic is quite feasible; still, the more realistic recognition systems will require Pashto-specific datasets.

3.2.2 Recent Advancements and Techniques

Despite these challenges, outstanding achievements have been made in Pashto handwriting recognition. The researchers have only recently started using deep learning techniques like CNNs which were used with success on Arabic and Persian handwriting. These models learn features automatically from the data and are therefore eminently suited to deal with cursive scripts like Pashto, which is diverse and complex due to [55, 56].

Besides, transfer learning-like techniques enable the adaptation to Pashto of models trained on larger datasets in similar languages, such as Arabic; thus, the lack of large Pashto-specific datasets will be overcome. That would make use of common characteristics of the script to improve performance without needing a whole new dataset for every language.

Due to scarce resources, however, most of these researches are at their preliminary stages of development compared to other Arabic and Persian handwriting recognitions. Current research is, therefore, carried out for the development of new datasets, improving segmentation algorithms for cursive scripts, and refining deep learning architecture that could handle the unique features of Pashto.

3.2.3 The Importance of Pashto Handwriting Recognition

Considering that Pashto is one of the official languages of Afghanistan and its millions of users in Pakistan, the development of a robust Pashto handwriting recognition system remains paramount. Such systems can find applications in document digitization, automated form processing, and preservation of historical text. Such technologies will help government institutions and other educational departments where almost all documentation still relies on handwritten or printed material in the aforementioned regions[57].

Furthermore, the recognition of Pashto handwriting would also develop tools for multilingual text processing in Arabic-script-based languages so that these languages can be integrated into modern digital systems. With the improvement of Pashto handwriting recognition, the research contributes to the development of other related languages like Urdu and Farsi since there is a shared feature in the script.

4 | Conclusion

Urdu handwriting recognition, particularly in the Nastaliq script, is a difficult but important topic of research in artificial intelligence and pattern recognition. Urdu is substantially more difficult to recognize than other scripts such as English or Arabic, despite certain structural similarities. While deep learning in general, and CNNs in particular, has shown promising results that improve recognition accuracy, the improvement of the state-of-the-art performance still suffers due to a lack of large and diverse data sets. This survey provided an overview of the existing challenges and the developments in Urdu handwriting recognition, specifically leveraging the few datasets UNHD, UHTD, and UHCD, which have provided relevant albeit limited resources for model training. Though there has been some success in offline handwriting recognition, the complexity of the script and variability in individual writing styles require more sophisticated approaches and better datasets to reach the accuracy level seen for other languages. Larger and more diverse datasets need to be developed, and techniques such as transfer learning from languages like Arabic should be exploited in the future to overcome dataset limitations. This would also be researched for more advanced deep learning architecture and segmentation techniques peculiar to the specific Urdu script. In addition, overcoming these challenges will further improve digital documentation and document processing in the Urdu-speaking regions, contributing to the improvement of Arabic-script-based handwriting recognition systems as a whole.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

All authors contributed equally to this work.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Lin C, Ahmad A, Qu R, et al (2024) A Handwriting Recognition System With WiFi. IEEE Trans Mob Comput 23:3391–3409. https://doi.org/10.1109/TMC.2023.3279608

[2] Hamida S, Cherradi B, El Gannour O, et al (2023) Cursive Arabic handwritten word recognition system using majority voting and k-NN for feature descriptor selection. Multimed Tools Appl 82:40657–40681. https://doi.org/10.1007/S11042-023-15167-6

[3] Al-wajih E, Ghazali R (2023) Threshold center-symmetric local binary convolutional neural networks for bilingual handwritten digit recognition. Knowl Based Syst 259:110079. https://doi.org/10.1016/J.KNOSYS.2022.110079

[4] Basu S, Das N, Sarkar R, et al (2010) A novel framework for automatic sorting of postal documents with multi-script address blocks. Pattern Recognit 43:3507–3521. https://doi.org/10.1016/J.PATCOG.2010.05.018

[5] Zhao A, Li J (2023) A significantly enhanced neural network for handwriting assessment in Parkinson's disease detection. Multimed Tools Appl 82:38297–38317. https://doi.org/10.1007/S11042-023-14647-Z

[6] Dargan S, Kumar M, Mittal A, Kumar K (2024) Handwriting-based gender classification using machine learning techniques. Multimed Tools Appl 83:19871–19895. https://doi.org/10.1007/S11042-023-16354-1

[7] Huang Z, Shivakumara P, Kaljahi MA, et al (2023) Writer age estimation through handwriting. Multimed Tools Appl 82:16033–16055. https://doi.org/10.1007/S11042-022-13840-W

[8] Singh S, Sharma A, Chhabra I (2017) A dominant points-based feature extraction approach to recognize online handwritten strokes. International Journal on Document Analysis and Recognition 20:37–58. https://doi.org/10.1007/S10032-016-0279-X

[9] Singh PK, Chatterjee I, Sarkar R, et al (2021) A new feature extraction approach for script invariant handwritten numeral recognition. Expert Syst 38:e12699. https://doi.org/10.1111/EXSY.12699

[10] Singh S, Sharma A (2019) Online Handwritten Gurmukhi Words Recognition. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 18:. https://doi.org/10.1145/3282441

[11] Singh S, Sharma A, Chauhan VK (2021) Online handwritten Gurmukhi word recognition using fine-tuned Deep Convolutional Neural Network on offline features. Machine Learning with Applications 5:100037. https://doi.org/10.1016/J.MLWA.2021.100037

[12] Singh PK, Chatterjee I, Sarkar R, et al (2021) A new feature extraction approach for script invariant handwritten numeral recognition. Expert Syst 38:e12699. https://doi.org/10.1111/EXSY.12699

[13] Roy S, Das N, Kundu M, Nasipuri M (2017) Handwritten isolated Bangla compound character recognition: A new benchmark using a novel deep learning approach. Pattern Recognit Lett 90:15–21. https://doi.org/10.1016/J.PATREC.2017.03.004

[14] Das N, Reddy JM, Sarkar R, et al (2012) A statistical–topological feature combination for recognition of handwritten numerals. Appl Soft Comput 12:2486–2495. https://doi.org/10.1016/J.ASOC.2012.03.039

[15] Das N, Sarkar R, Basu S, et al (2012) A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application. Appl Soft Comput 12:1592–1606. https://doi.org/10.1016/J.ASOC.2011.11.030

[16] Mukhoti J, Dutta S, Sarkar R (2020) Handwritten Digit Classification in Bangla and Hindi Using Deep Learning. Applied Artificial Intelligence 34:1074–1099. https://doi.org/10.1080/08839514.2020.1804228

[17] Hijam D, Saharia S (2022) On developing complete character set Meitei Mayek handwritten character database. Visual Computer 38:525–539. https://doi.org/10.1007/S00371-020-02032-Y

[18] Porwal U, Fornés A, Shafait F (2022) Advances in handwriting recognition. International Journal on Document Analysis and Recognition 25:241–243. https://doi.org/10.1007/S10032-022-00421-8

[19] Inunganbi S (2024) A systematic review on handwritten document analysis and recognition. Multimed Tools Appl 83:5387–5413. https://doi.org/10.1007/S11042-023-15326-9

[20] ul Sehr Zia N, Naeem MF, Raza SMK, et al (2022) A convolutional recursive deep architecture for unconstrained Urdu handwriting recognition. Neural Comput Appl 34:1635–1648. https://doi.org/10.1007/S00521-021-06498-2

[21] ul Sehr Zia N, Naeem MF, Raza SMK, et al (2022) A convolutional recursive deep architecture for unconstrained Urdu handwriting recognition. Neural Comput Appl 34:1635–1648. https://doi.org/10.1007/S00521-021-06498-2

[22] Cheema MDA, Shaiq MD, Mirza F, et al (2024) Adapting multilingual vision language transformers for low-resource Urdu optical character recognition (OCR). PeerJ Comput Sci 10:1–24. https://doi.org/10.7717/PEERJ-CS.1964

[23] Shaik MA, Abdul W (2023) Optical Character Recognition for Urdu Text: A Review of Techniques, Challenges, and Future Directions. Proceedings of IEEE 2023 5th International Conference on Advances in Electronics, Computers and Communications, ICAECC 2023. https://doi.org/10.1109/ICAECC59324.2023.10560177

[24] Nabi ST, Singh P, Kumar M (2023) Writer Identification from Offline Handwriting Images in Urdu Script with Dense-Net: A Deep Learning Approach. 2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023. https://doi.org/10.1109/ICCCNT56998.2023.10307034

[25] [Nabi ST, Kumar M, Singh P (2024) DeepNet-WI: a deep-net model for offline Urdu writer identification. Evolving Systems 15:759–769. https://doi.org/10.1007/S12530-023-09504-1

[26] Ahmed S Bin, Naz S, Swati S, Razzak MI (2019) Handwritten Urdu character recognition using one-dimensional BLSTM classifier. Neural Comput Appl 31:1143–1151. https://doi.org/10.1007/S00521-017-3146-X

[27] Bin Ahmed S, Naz S, Swati S, et al (2017) UCOM offline dataset-an Urdu handwritten dataset generation. Volume 14, Issue 2, Pages 239 - 245 14:239–245

[28] Rafique A, Ishtiaq M (2022) UOHTD: Urdu Offline Handwritten Text Dataset. In: Porwal U, Fornés A, Shafait F (eds) Frontiers in Handwriting Recognition. Springer International Publishing, Cham, pp 498–511

[29] Nanehkaran YA, Chen J, Salimi S, Zhang D (2021) A pragmatic convolutional bagging ensemble learning for recognition of Farsi handwritten digits. Journal of Supercomputing 77:13474–13493. https://doi.org/10.1007/S11227-021-03822-4

[30] Sagheer MW, He CL, Nobile N, Suen CY (2009) A new large urdu database for off-line handwriting recognition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5716 LNCS:538–546. https://doi.org/10.1007/978-3-642-04146-4_58

[31] Husnain M, Saad Missen MM, Mumtaz S, et al (2020) Urdu handwritten text recognition: a survey. IET Image Process 14:2291–2300. https://doi.org/https://doi.org/10.1049/iet-ipr.2019.0401

[32] Rasheed A, Ali N, Zafar B, et al (2022) Handwritten Urdu Characters and Digits Recognition Using Transfer Learning and Augmentation With AlexNet. IEEE Access 10:102629–102645. https://doi.org/10.1109/ACCESS.2022.3208959

[33] Hijam D, Saharia S (2022) On developing complete character set Meitei Mayek handwritten character database. Vis Comput 38:525–539. https://doi.org/10.1007/s00371-020-02032-y

[34] Misgar MM, Mushtaq F, Khurana SS, Kumar M (2023) Recognition of offline handwritten Urdu characters using RNN and LSTM models. Multimed Tools Appl 82:2053–2076. https://doi.org/10.1007/S11042-022-13320-1

[35] Mushtaq F, Misgar MM, Kumar M, Khurana SS (2021) UrduDeepNet: offline handwritten Urdu character recognition using deep neural network. Neural Comput Appl 33:15229–15252. https://doi.org/10.1007/S00521-021-06144-X

[36] Naz S, Umar AI, Ahmad R, et al (2017) Urdu Nastaliq recognition using convolutional–recursive deep learning. Neurocomputing 243:80–87. https://doi.org/10.1016/J.NEUCOM.2017.02.081

[37] Zia S, Azhar M, Lee B, et al (2023) Recognition of printed Urdu script in Nastaleeq font by using CNN-BiGRU-GRU Based Encoder-Decoder Framework. Intelligent Systems with Applications 18:200194. https://doi.org/10.1016/J.ISWA.2023.200194

[38] Riaz N, Arbab H, Maqsood A, et al (2022) Conv-transformer architecture for unconstrained off-line Urdu handwriting recognition. International Journal on Document Analysis and Recognition 25:373–384. https://doi.org/10.1007/S10032-022-00416-5

[39] Naz S, Umar AI, Ahmed R, et al (2016) Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks. Springerplus 5:1–16. https://doi.org/10.1186/S40064-016-3442-4

[40] Naz S, Umar AI, Ahmad R, et al (2016) Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks. Neurocomputing 177:228–241. https://doi.org/10.1016/J.NEUCOM.2015.11.030

[41] Naz S, Umar AI, Ahmad R, et al (2017) Urdu Nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features. Neural Comput Appl 28:219–231. https://doi.org/10.1007/S00521-015-2051-4

[42] Owens J (2013) The Oxford handbook of Arabic linguistics. Oxford University Press

[43] Ahmed R, Dashtipour K, Gogate M, et al (2020) Offline arabic handwriting recognition using deep machine learning: A review of recent advances. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11691 LNAI:457–468. https://doi.org/10.1007/978-3-030-39431-8_44

[44] Alrobah N, Albahli S (2022) Arabic Handwritten Recognition Using Deep Learning: A Survey. Arab J Sci Eng 47:9943–9963. https://doi.org/10.1007/s13369-021-06363-3

[45] Alghyaline S (2023) Arabic Optical Character Recognition: A Review. CMES - Computer Modeling in Engineering and Sciences 135:1825–1861

[46] Jemni SK, Kessentini Y, Kanoun S (2019) Out of vocabulary word detection and recovery in Arabic handwritten text recognition. Pattern Recognit 93:507–520. https://doi.org/10.1016/j.patcog.2019.05.003

[47] Liu L, Ouyang W, Wang X, et al (2020) Deep Learning for Generic Object Detection: A Survey. Int J Comput Vis 128:261–318. https://doi.org/10.1007/S11263-019-01247-4

[48] Guptha NS, Balamurugan V, Megharaj G, et al (2022) Cross lingual handwritten character recognition using long short term memory network with aid of elephant herding optimization algorithm. Pattern Recognit Lett 159:16–22. https://doi.org/10.1016/j.patrec.2022.04.038

[49] Ali AAA, Mallaiah S (2022) Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout. Journal of King Saud University - Computer and Information Sciences 34:3294–3300. https://doi.org/10.1016/J.JKSUCI.2021.01.012

[50] Nayef BH, Abdullah SNHS, Sulaiman R, Alyasseri ZAA (2022) Optimized leaky ReLU for handwritten Arabic character recognition using convolution neural networks. Multimed Tools Appl 81:2065–2094. https://doi.org/10.1007/s11042-021-11593-6

[51] Elkhayati M, Elkettani Y (2022) UnCNN: A New Directed CNN Model for Isolated Arabic Handwritten Characters Recognition. Arab J Sci Eng 47:10667–10688. https://doi.org/10.1007/s13369-022-06652-5

[52] Khaliq F, Shabir M, Khan I, et al (2023) Pashto Handwritten Invariant Character Trajectory Prediction Using a Customized Deep Learning Technique. Sensors 2023, Vol 23, Page 6060 23:6060. https://doi.org/10.3390/S23136060

[53] Siddhu MK, Yaakob SN (2019) Deep learning applied to arabic and latin scripts: A review. Volume 8, Issue 11, Pages 1510 - 1521 8:1510–1521

[54] Khan M, Rahman TU, Sher M, et al (2024) Flexible Ionic Conductive Hydrogels with Wrinkled Texture for Flexible Strain Transducer with Language Identifying Diversity. Chemistry of Materials 36:4703–4713.

[55] Khaliq F, Shabir M, Khan I, et al (2023) Pashto Handwritten Invariant Character Trajectory Prediction Using a Customized Deep Learning Technique. Sensors 2023, Vol 23, Page 6060 23:6060. https://doi.org/10.3390/S23136060

[56] Rehman MZ, Nawi NM, Arshad M, Khan A (2021) Recognition of Cursive Pashto Optical Digits and Characters with Trio Deep Learning Neural Network Models. Electronics 2021, Vol 10, Page 2508 10:2508. https://doi.org/10.3390/ELECTRONICS10202508

[57] Ahmad R, Naz S, Razzak I (2021) Efficient skew detection and correction in scanned document images through clustering of probabilistic hough transforms. Pattern Recognit Lett 152:93–99. https://doi.org/10.1016/J.PATREC.2021.09.014