

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/382255843>

Optical Character Recognition of Balochi Script

Article in International Journal of Scientific Research in Computer Science Engineering and Information Technology · July 2024

DOI: 10.32628/CSEIT241046

CITATION

1

READS

42

4 authors, including:



Dil Nawaz Hakro
University of Sindh

67 PUBLICATIONS 313 CITATIONS

[SEE PROFILE](#)



Abdul Majid
Ocean University of China

7 PUBLICATIONS 4 CITATIONS

[SEE PROFILE](#)

Optical Character Recognition of Balochi Script

¹Muhammad Mazhar, ¹Qinbo, ²Dil Nawaz Hakro, ¹Abdul Majid

¹Department of Computer Science and Technology, Faculty of Information Science and Technology,
Ocean University of China

²Faculty of Engineering and Technology (FET) University of Sindh, Jamshoro, Pakistan

Corresponding Author: Muhammad Mazhar, muhammadmazhar4097@gmail.com

ARTICLE INFO

Article History:

Accepted : 26 June 2024

Published: 14 July 2024

Publication Issue

Volume 10, Issue 4

July-August-2024

Page Number

115-124

ABSTRACT

Optical Character Recognition is considered one of the fastest methods of data entry. OCR converts the text image representation of x and y coordinates representing pixel information to be converted into text data in a particular language. OCR as a field of pattern recognition and document image understanding, OCR requires a challenging job once a different language text is available on the image. Difference in language script will pose different challenges for OCR which requires entirely different approaches and algorithms. Latin scripts require a different approach whereas the Balochi adopted language scripts require a different approach. In this regard, various solutions have been proposed for different languages. Segmentation is considered one of the important tasks in the process of OCR. A good segmentation will definitely increase the accuracy of an OCR. Segmentation includes the segmentation of text lines from text images which are further divided into words. These segmented words are further divided into characters which are to be recognized. A single segmentation algorithm to segment various scripts of the languages is proposed in this study which checks the script and then segments the text image for the further processing in OCR. The proposed generalized algorithm will check the style, direction and other properties of the script and then adopts the segmentation process to segment text lines, words and characters of the language. The proposed algorithm segments more than ten languages of three scripts and segments for their OCRs. These images can be further processed for feature extraction and classification further. The process of OCR for selected languages will be made easier to recognize. Multiple scripts, languages and images were experimented, and the proposed algorithm successfully segmented 42,833 images of text line, words and character image. The algorithm provides 97% accuracy while segmenting these images and can be extended to further languages as well as scripts .

Keywords : Optical Character Recognition of Balochi Script

I. INTRODUCTION

Balochi language is spoken in southwestern Pakistan, especially in the province of Baluchistan and Sindh by a large number of people. The language is also a great source of communication for the people who settled in the northeastern regions of Khorasan and Sistan Baluchistan which is the second largest province in Iran. It is also spoken by smaller communities settled in Afghanistan, Oman, United Arab Emirates, Turkmenistan, India, East Africa, and Bahrain. The Balochi script is cursive and written from right to left. It inherits some of its characteristics from Arabic, Farsi, and Urdu scripts, additionally, it has a larger number of characters such as differentiating characters, dot characters, range of location, and dot orientations that have been minimized over time due to advancement in Balochi script. Balochi or Balochi is an Iranian language spoken primarily in the Baluchistan region divided between Pakistan, Iran and Afghanistan. Balochi belongs to the Northwestern Iranian linguistic classification. It is spoken by 3 to 5 million people. In addition to Pakistan, Iran and Afghanistan, it is also spoken in Oman, the Arab states of the Persian Gulf, Turkmenistan, and East Africa and in diaspora communities in other parts of the world. The Balochi Language is made up of 40 alphabets and 7 special characters. The Balochi script also consists of seven special characters, these special characters contain dots or a diacritic above and below the characters. A typical OCR system may contain the following steps shown in Figure 1.

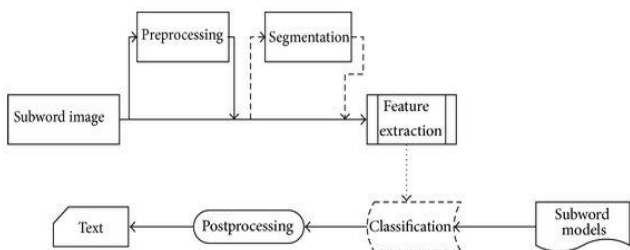


Figure 1: Elements of typical Optical Character Recognition

No.	Original text	Edges identification	Segmented Characters
1.	سومرو		
2.	پٹاٹ		

Figure 2: Horizontal projection profile segmentation of Sindhi Characters [11]

OCR of balochi script can be challenging task due to the complexity of the script and the lack of standardization in its writing however, there have been some efforts to develop OCR systems for balochi script. OCR of balochi script is still a developing field and more research is needed to improve the accuracy and efficiency of OCR systems for this script. Optical character recognition is a subarea of AI that converts scanned text images into an editable document. The researchers proposed various text recognition techniques to identify cursive and connected scripts written from left to right but their correct recognition is still a challenging problem for the visual methods. The Balochi language is one of them spoken by a significant part of the world population and no properly research conducted on the character recognition this regional language of Pakistan. Before detaining the OCR works, the peculiarities of the Balochi-like scripts are outlined, which are followed by the presentation of the available text image databases. For the sake of clarity, the various attempts are grouped into three parts, namely: (a) printed, (b) handwritten, and (c) online character recognition. We aim to develop an OCR system for balochi script to improve the accessibility of balochi language material.





Figure 2a: Segmentation of Balochi text image (a)
Original image (b) Segmented Text Lines



Figure 2b: Segmentation of Balochi text image (a)
Original image (b) Segmented Text Lines

II. LITERATURE REVIEW

Researchers discussed various approaches and techniques in their research for the recognition of regional language characters and performed segmentation on these characters. There have been

some efforts to develop OCR system for other language that use the Arabic script, such as urdu and persian. However balochi script has its own unique characteristics, such as the use of additional diacritic marks and ligatures, which make it more challenging to recognize. Some research has been conducted on the recognition of balochi script using ruled based and machine learning techniques. However, there is still a needed for more research to improve the accuracy and efficiency of OCR systems for Balochi script.

Ghulam et al. [1] for Balochi script recognition for non-cursive characters a convolutional neural network based model is used. This model optimized small VGGNet model and achieved exceptional precision and speed over the state of the art methods of machine learning We tested the method with randomly collected images and the model correctly recognized the characters with 96% accuracy.

S.M Lodhi et al. [2] used the Fourier descriptor technique for the recognition of Urdu characters. The descriptor characterized the shape and features even in the presence of noise. The descriptor also performed the scaling and interpretation even if the shape and position of the characters are changed.

Lorigo and Govindaraju [3] used a combination of different methods based on artificial neural networks. They used Hidden Markov model and contour-based approach for the recognition of Arabic handwritten script. Arabic script is written from right to left and characters are joined in a machine-readable format. They also discussed the representation of Arabic letters, words, and analyzed handwritten methods. Mohammad et al. [4] also explored a segmentation and recognition technique for Arabic text and using a contour-based approach that found out edges and the region of interest for the sub-words and claimed improvements over the finding of the skeleton of the word.

Solimanpour et al. [4] explored the contour-based method for the recognition of Farsi character recognition. They prepared the Farsi language dataset comprised of characters, dates, and numerical strings. They also created the Farsi dataset for further research. Experiments were performed on the dataset and claimed better recognition results on Farsi characters and digits.

Shamsher et al. [5] explored the supervised learning to train a feed-forward neural network, later used the network for the identification of non-cursive characters. The proposed technique performed better during the training phase and claimed better recognition results.

Sattar et al. [6] used a Markov model (MM) for the recognition of the Urdu alphabets. They selected the full paragraph rather than isolated characters. The Markov model process a word as a chain of individual characters and considered sentence another chain of words. They extracted the features of each character and calculated the probability of recognizing each character.

Akhbari et al. [7] presented a projection-based technique in which horizontal and vertical (x, y) histogram projection is applied to each line. The method detected the words and characters to divide the text lines. They proposed three steps to perform division such as segmentation of text lines from Arabic script images, divided the lines into words using blank spaces present between the words and finally used vertical projection technique for the segmentation of connected words.

Shaikh et al. [8] proposed a technique for the extraction of characters from the sub-words of cursive text. The algorithm used height profile vector (HPV) in which the difference between the first most pixel in each column of the sub-word and the baseline pixel for the segmentation of thinned stroke sub-words. The

method helped them to find the location of the segmented points of the characters.

Alaei et al. [9] explored a method for Persian handwritten character recognition. The shapes are categorized into 8 different shapes out of a total of 32 Persian characters using a bitmap technique. In the bitmap technique, each character is recognized by a sliding window of size 7×7 to extract features. Finally, the support vector machine (SVM) algorithm is used for the classification of the text.

Taha et al. [10] proposed a method for the Arabic printed text character recognition. The proposed method consists of the following steps; image acquisition and preprocessing, segmentation of characters, feature extraction, and finally, character recognition.

D.N Hakro et al. [11] used optical character recognition technique to recognize Sindhi characters. The proposed method consists of basic image preprocessing steps to remove noise present in the target images and used a template matching technique for the character recognition. The segmentation of isolated Sindhi characters is shown in Figure 3.


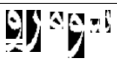

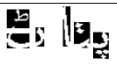
No.	Original text	Edges identification	Segmented Characters
1.	سومرو		
2.	پٹاٹ		

Figure 3 : Horizontal projection profile segmentation of Sindhi Characters [11]

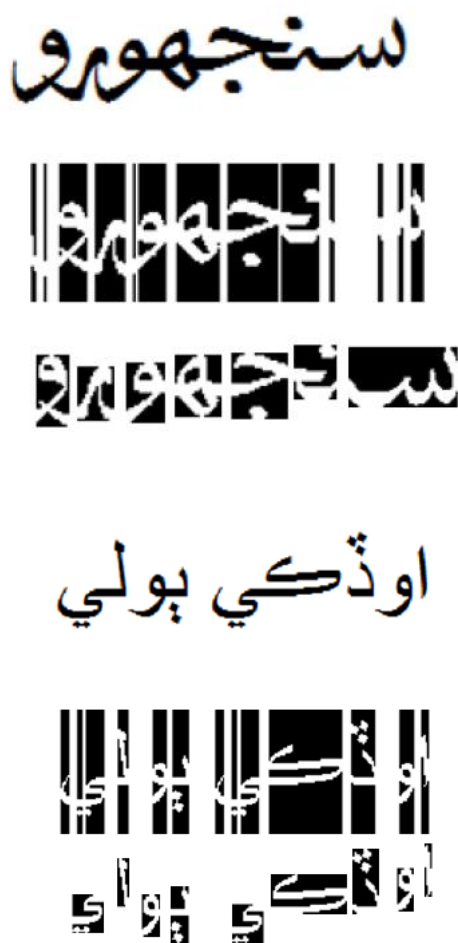


Figure 4: Horizontal projection profile segmentation of Sindhi characters [11]

Ahmed et al. [12] compared three different classifiers for the recognition of Urdu printed characters. The method consisted of a scale-invariant feature transform (SIFT), long shortterm memory (LSTM) and Markov model. They analyzed that LSTM outperformed the other two baseline methods for character recognition.

Srivastava et al [13] optical character recognition (OCR) techniques, including traditional methods such as template matching, feature extraction-based approaches, and recent advancements in deep learning-based OCR models. The review highlights the strengths and limitations of different OCR techniques, explores their applications in document analysis, text recognition, and handwriting recognition, and

identifies areas for further research and improvement in OCR technology.

Alghyaline et al [14] Arabic Optical Character Recognition (OCR) highlights the significant advancements in Arabic OCR techniques, including feature extraction methods, machine learning algorithms, and deep learning approaches, to address the challenges posed by the complex Arabic script. The review emphasizes the importance of accurate preprocessing techniques, segmentation algorithms, and language-specific considerations in achieving high recognition accuracy for Arabic OCR systems, showcasing the progress made in the field and suggesting future research directions.

Siddiqu et al [15] Baseline Isolated Printed Text Image Database for Pashto Script Recognition. In the context of Pashto script recognition, the significance of the Baseline Isolated Printed Text Image Database, which addresses the need for a dedicated dataset for Pashto script OCR. The review showcases the importance of the database in enabling research and development of Pashto OCR techniques, including feature extraction methods, segmentation algorithms, and machine learning models. Furthermore, the review emphasizes the value of the Baseline Isolated Printed Text Image Database in fostering advancements in Pashto script recognition, facilitating applications such as document digitization, text analysis, and language processing.

Sanjrani et al [16] Multilingual Optical Character Recognition (OCR) systems tailored for the regional languages in Baluchistan, such as Balochi, Brahui, and others. The review identifies the challenges specific to these languages, including complex character shapes, ligatures, and variations in writing styles, and explores various research efforts in dataset creation, feature extraction techniques, segmentation algorithms, and machine learning models. Furthermore, the review emphasizes the need for language-specific preprocessing methods, performance evaluation

metrics, and language-specific OCR techniques to bridge the research gap in this domain and enable the accurate recognition and processing of regional language texts in Baluchistan.

Abir et al [17] Bengali Optical Character Recognition (OCR) system, including the complexities of Bengali script, lack of suitable datasets, inadequate preprocessing techniques, and the need for robust recognition algorithms, highlights the research gaps in this field.

Pulabaigari et al. [18] an efficient multi-lingual OCR system for Indian languages using the Bharati script, focuses on the challenges related to script variations, language-specific features, dataset availability, preprocessing techniques, and recognition algorithms, highlighting the advancements made and identifying potential research gaps.

These studies collectively highlight the efforts in developing OCR systems specifically for the Balochi script. They employ various techniques, including feature extraction, machine learning algorithms, and deep learning models, to accurately recognize Balochi characters and text. The studies demonstrate promising results and emphasize the importance of considering the unique characteristics and complexities of the Balochi script for achieving accurate OCR in Balochi.

2- Research Gap

The research gap lies in the absence of comprehensive datasets, suitable preprocessing techniques, and robust recognition algorithms specifically tailored for the complexities of the Balochi script, hindering the development of accurate and effective OCR systems for Balochi text.

3- Motivation

The development of an OCR system for Balochi script will have several benefits. It will improve the

accessibility of Balochi language materials, making it easier for people to read and understand them. It will also facilitate the digitization of Balochi language materials, which will help to preserve and promote the language. Additionally, it will contribute to the field of OCR research by addressing the challenging associated with recognizing Balochi script.

4- Problem statement

Optical Character Recognition (OCR) of the Balochi script presents a challenging task due to the lack of comprehensive resources and established techniques specifically tailored for this script. Balochi, a widely spoken language with its unique writing system, is not well-represented in existing OCR systems, hindering its utilization in various domains such as document digitization, language processing, and information retrieval.

The absence of an accurate and reliable OCR solution for the Balochi script impedes the automation of tasks that require the extraction and analysis of text from printed or handwritten documents. Existing OCR systems primarily focus on widely-used scripts, resulting in limited support for Balochi script recognition, which often leads to significant errors and inaccuracies.

To address this problem, this research aims to develop an effective OCR system specifically designed for the Balochi script. This entails the creation of a comprehensive dataset of labeled Balochi script samples, considering the variations in handwriting styles, ligatures, and diacritic marks commonly found in the Balochi text. Furthermore, the research will explore suitable preprocessing techniques to enhance the quality of the input images and address challenges like noise, skew, and variations in font styles.

The key challenges to be addressed in this research include the development of robust feature extraction methods for Balochi characters, the design and training

of an OCR model capable of accurately recognizing Balochi script, and the optimization of the model to achieve high accuracy and efficiency. Additionally, the limited availability of annotated data for Balochi script poses a challenge, necessitating the exploration of data augmentation techniques or transfer learning from related scripts.

The successful implementation of an OCR system for Balochi script will empower various applications, including digitizing historical documents, enabling efficient information retrieval, facilitating language translation, and supporting Balochi language preservation and analysis.

5- Research Aim

The aim of this research is to develop an OCR system for Balochi script that can accurately and efficiently recognize Balochi characters. The system will be based on a combination of ruled-based and deep learning techniques, and it will be evaluated using metrics such as accuracy, precision, and recall. The research will contribute to the field of OCR research by addressing the challenges associated with recognizing Balochi script and improving the accessibility of Balochi language materials.

6- Research Objective

In this research, we will use various tools and technologies to develop the OCR system for Balochi script. The ruled-based approach will involve using regular expressions and pattern-matching algorithms to identify and classify characters in the next. The deep learning approach will involve using Python programming language and deep learning frameworks such as Tensor Flow and Keras to train and evaluate the CNN model. We will also use image processing libraries such as OpenCV to reprocess the Balochi text images.

7- Research Scope / Limitation

The scope of optical character recognition (OCR) of Balochi script is to accurately recognize and convert printed, handwritten and online character recognition

of Balochi text into digital format. This can be useful for digitizing historical documents, newspapers, and books written in Balochi language. OCR technology can also be used to create searchable databases of Balochi texts and to enable machine translation of Balochi text. However, OCR of Balochi script faces several limitations. Balochi script is not standardized, and there are several variations of the script, which make it difficult for OCR software to accurately recognize the characters. Additionally, Balochi script uses ligature and diacritical marks, which can be challenging for OCR software to distinguish. Handwritten Balochi text can also be difficult to recognize due to variations in handwriting styles. Furthermore, the lack of digital resources and training data for Balochi OCR can hinder the development of accurate OCR software. This can also limit the ability to develop language models and machine learning algorithms for Balochi OCR.

8- Research plan/ Methodology (Schedule/Phasing)

As of my knowledge cutoff in September 2021, there wasn't a widely known or widely implemented Optical Character Recognition (OCR) system specifically designed for the Balochi script. OCR technology is primarily developed and implemented for widely used scripts like Latin, Cyrillic, Chinese, Japanese, and Korean.

In this research, we will use a combination of ruled-based and deep learning techniques to develop an OCR system for Balochi script and use them to identify and classify characters in the next. The deep learning approach will involve training a convolutional neural network (CNN) on a large dataset of Balochi text images and using it to classify new images of Balochi text. We will evaluate the performance of the OCR system using metrics such as accuracy, precision, and recall.

Developing an Optical Character Recognition (OCR) system for the Balochi script would require a well-

structured research plan. Here is a generalized outline of a research plan that could be followed:

1. Literature Review:

- Conduct a comprehensive review of existing literature on OCR technologies, methodologies, and algorithms.
- Identify relevant research papers, articles, and resources on OCR systems for similar scripts/languages.

2. Data Collection:

- Collect a diverse dataset of Balochi script samples, including different fonts, styles, sizes, and image qualities.
- Ensure the dataset represents a wide range of variations and challenges present in Balochi script.

3. Data Preprocessing:

- Apply necessary preprocessing techniques to enhance the quality of the collected dataset.
- Explore methods for noise reduction, normalization, and image enhancement specific to Balochi script.

4. Character Segmentation:

- Investigate and develop algorithms for accurately segmenting individual characters in the Balochi script.
- Address challenges such as connected characters or complex ligatures.

5. Feature Extraction:

- Explore various feature extraction techniques suitable for Balochi script recognition.
- Investigate both structural and appearance-based features that can capture the unique characteristics of Balochi script.

6. Training and Model Development:

- Utilize machine learning or deep learning algorithms to train a model for Balochi script recognition.
- Design appropriate architectures and optimize hyperparameters based on the dataset and research goals.

7. Evaluation Metrics:

- Define evaluation metrics to assess the performance of the OCR system, such as accuracy, precision, recall, and F1 score.
- Establish a testing framework using a separate dataset to evaluate the system's performance.

8. Performance Evaluation:

- Evaluate the trained OCR model using the defined metrics and testing framework.
- Analyze the strengths, weaknesses, and limitations of the system.
- Identify areas of improvement and potential sources of errors.

9. Iterative Refinement:

- Refine the OCR system based on the evaluation results.
- Explore techniques to address specific challenges or limitations encountered during the evaluation.

10. Comparative Analysis:

- Compare the performance of the developed OCR system with existing OCR systems for other languages or scripts.
- Identify areas where the Balochi OCR system excels or requires further improvement.

11. Application and Integration:

- Explore potential applications of the Balochi OCR system in real-world scenarios.
- Investigate integration options with existing software or platforms to facilitate Balochi script recognition.

12. Documentation and Publication:

- Document the research findings, methodologies, and techniques used in developing the Balochi OCR system.
- Publish research papers, articles, or technical reports to contribute to the OCR research community.

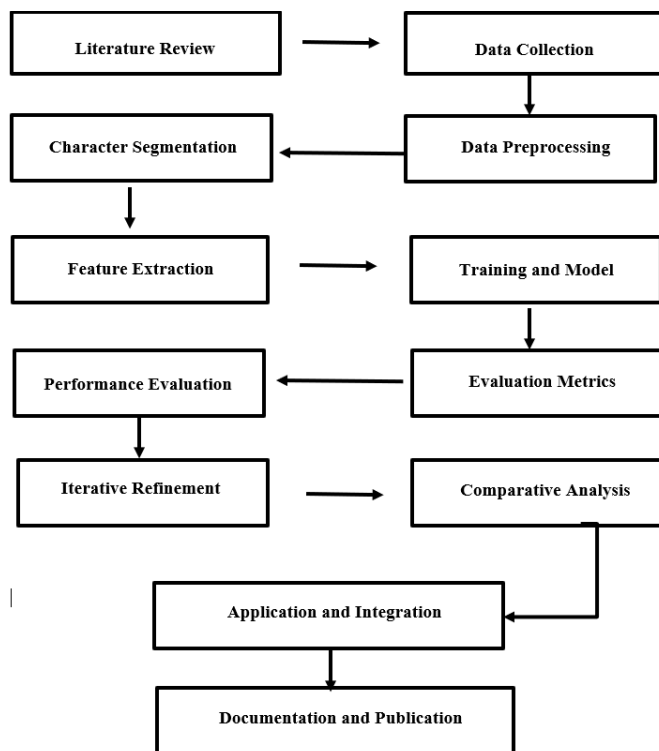


Figure 5: Proposed Research Framework for Balochi OCR

III. REFERENCES

- [1]. Naseer, G. J., Basit, A., Ali, I., & Iqbal, A. (2020). Balochi Non-Cursive Isolated Character Recognition using Deep Neural Network. *International Journal of Advanced Computer Science and Applications*, 11(4).
- [2]. S. Lodhi and M. Matin, "Urdu character recognition using Fourier descriptors for optical networks," in *Photonic Devices and Algorithms for Computing VII*, vol. 5907. International Society for Optics and Photonics, 2005, p. 59070O.
- [3]. L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 28, no. 5, pp. 712–724, 2006.
- [4]. F. Solimanpour, J. Sadri, and C. Y. Suen, "Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language," 2006.
- [5]. I. Shamsheer, Z. Ahmad, J. K. Orakzai, and A. Adnan, "Ocr for printed urdu script using feed-forward neural network," in *Proceedings of World Academy of Science, Engineering and Technology*, vol. 23. Citeseer, 2007, pp. 172–175.
- [6]. S. A. Sattar, S. Haque, M. K. Pathan, and Q. Gee, "Implementation challenges for Nastaliq character recognition," in *International Multi-Topic Conference*. Springer, 2008, pp. 279–285.
- [7]. Z. Al Aghbari and S. Brook, "Hah manuscripts: A holistic paradigm for classifying and retrieving historical Arabic handwritten documents," *Expert Systems with Applications*, vol. 36, no. 8, pp. 10 942–10 951, 2009.
- [8]. N. A. Shaikh, G. A. Mallah, and Z. A. Shaikh, "Character segmentation of Sindhi, an Arabic style scripting language, using height profile vector," *Australian Journal of Basic and Applied Sciences*, vol. 3, no. 4, pp. 4160–4169, 2009.
- [9]. A. Alaei, P. Nagabhushan, and U. Pal, "A new two-stage scheme for the recognition of Persian handwritten characters," in *2010 12th International Conference on Frontiers in Handwriting Recognition*. IEEE, 2010, pp. 130–135.
- [10]. S. Taha, Y. Babiker, and M. Abbas, "Optical character recognition of Arabic printed text," in *2012 IEEE Student Conference on Research and Development (SCORED)*. IEEE, 2012, pp. 235–240.
- [11]. D. N. Hakro, I. A. Ismaili, A. Z. Talib, Z. Bhatti, and G. N. Mojai, "Issues and challenges in Sindhi ocr," *Sindh University Research Journal (Science Series)*, vol. 46, no. 2, pp. 143–152, 2014.
- [12]. Ahmad, I., Wang, X., Li, R., & Rasheed, S. (2017). Offline Urdu Nastaleeq optical character recognition based on stacked denoising autoencoder. *China Communications*, 14(1), 146–157.

- [13]. Srivastava, S., Verma, A., & Sharma, S. (2022, February). Optical character recognition techniques: A review. In 2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-6). IEEE.
- [14]. Alghyline, S. Arabic Optical Character Recognition: A Review.
- [15]. Siddiqu, A., Basit, A., Noor, W., Khan, M. A., Kakar, M. S. H., & Khan, A. (2023). Baseline Isolated Printed Text Image Database for Pashto Script Recognition. *Intelligent Automation & Soft Computing*, 37(1).
- [16]. Sanjrani, A. A., Naveed, M. S., Sajid, M., Ahmed, A., Awan, S., & Jumani, A. K. (2020). Multilingual OCR systems for the regional languages in Balochistan. *Indian Journal of Science and Technology*, 13(21), 2157-2167.
- [17]. Abir, A. S. M., Rahman, S., Ellin, S., Farzana, M., Manik, M. H., & Rahman, C. R. (2020). Constraints in Developing a Complete Bengali Optical Character Recognition System. no. March.
- [18]. Pulabaigari, V. (2019, July). An Efficient Multi-Lingual Optical Character Recognition System for Indian Languages Through Use of Bharati Script. In *Document Analysis and Recognition: 4th Workshop, DAR 2018, Held in Conjunction with ICVGIP 2018, Hyderabad, India, December 18, 2018, Revised Selected Papers* (Vol. 1020, p. 74). Springer.