Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

PROGRESS AND INNOVATION IN OCR FOR GUJARATI NEWSPAPER TEXT RECOGNITION

Vidhdhi J Rughani[1]and Prof. (Dr.) Atul M. Gonsai[2]
[1] Department of Computer Science and Technology, Saurashtra University Rajkot, INDIA
[2] Department of Computer Science and Technology, Saurashtra University Rajkot, INDIA
vidhdhi.rajdev@gmail.com, atul.gosai@gmail.com

Abstract - Character recognition is a crucial technology in natural language processing (NLP) that translates printed text into machine-readable formats, aiding applications in document digitization, information retrieval, and more. This review paper digs into the important role of character recognition systems, emphasizing their usefulness for digitizing Gujarati newspapers. As an extensive repository of cultural and informational content for millions of Gujarati speakers, the digitization of these newspapers with OCR technology is crucial for preservation, accessibility, and distribution. However, the distinctive peculiarities of Gujarati script provide specific issues that demand customized OCR solutions. This research thoroughly reviews existing OCR techniques and methodologies, notably those focused on the Gujarati language, and show the limitations and obstacles faced in this sector. Through a comprehensive literature review, the paper analyzes the evolution and comparison of various character recognition systems, from traditional methods to advanced machine learning and deep learning approaches. It also identifies research gaps and recommends future research topics, seeking to enhance the development and accuracy of Gujarati OCR systems. By addressing these deficiencies, the article hopes to contribute significantly to the improvement of OCR technology for Gujarati, boosting its accessibility and usefulness for digital preservation and educational reasons. The systematic review encompasses the introduction, background, available literature, problems, methodology, system assessments, and future directions, concluding in a summary of significant findings and the vital need for continuous study in this subject..

Keyword- Deep Learning, Optical Character Recognition, Handwritten Character Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Template Matching, Gujarati Script Recognition.

I. INTRODUCTION

Character recognition is a key technique for converting printed text into machine-readable text. This introduction section presents an overview of character recognition in natural language processing [1], the importance of character recognition systems in the context of Gujarati Newspapers [2]. Moreover, the scope and objectives of this review paper are explained, and a brief review of the structure is added, which summarizes the background, literature review, objectives, and a brief overview of the structure.

Character Recognition in NLP is the initial factor that involves the method of digitizing printed text into machine-readable text by performing computational operations [3]. This technology has many applications in numerous areas such as document digitizing, information retrieval, etc. [4]. The digitization of character recognition enables text accessibility of printed paper, allowing experts, students, and scholars to access enormous repositories for various purposes.

The Gujarati newspaper is an essential source of information and cultural expression for millions of Gujarati language speakers [3]. Though there is a strong demand for Gujarati newspapers to be digitized as OCR technology makes valuable information accessible, preserved, and shared. A character recognition system customized to the difficulties of Gujarati script is required for precise transcription, making it easier to incorporate into digital archives for journalists [5], historians, and students. It is important in fields including education, media, government, historical research, and language technology. Thus, character recognition improves their usability, preservation, and accessibility for coming generations [6].

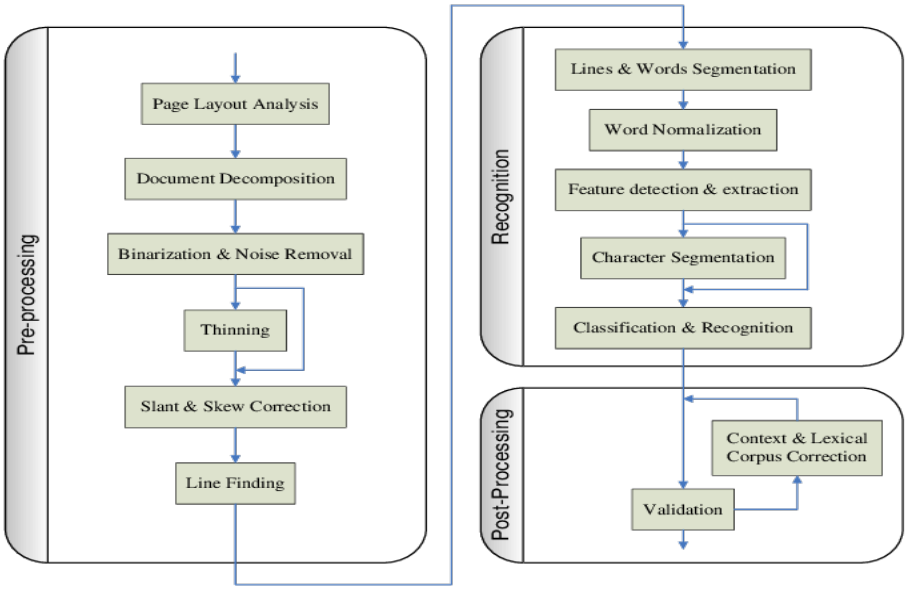

figure 1 A general framework for character recognition

The key objective of this review paper is to deeply evaluate existing OCR techniques and methodologies, the challenges, and limitations specifically tailored for the Gujarati language. The research goes beyond a basic analysis of current technologies to explore the complex challenges and constraints associated with Gujarati OCR systems in the present day. The purpose of the review article is to propose novel methodologies and provide guidance for addressing existing challenges and enhancing the advancement of Gujarati OCR technology. The paper aims to provide a significant contribution to the progress of OCR technology in the Gujarati language by implementing a comprehensive approach. It will improve accessibility, usability, and accuracy for digital document storage and educational initiatives in communities that speak Gujarati.

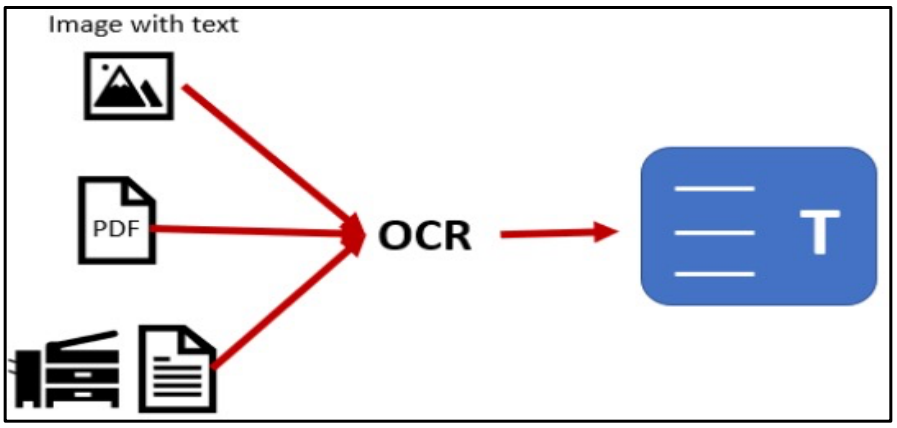

figure 2 Optical Character Recognition (OCR)

This review paper provides Section 1: Introduction, which summarizes OCR in natural language processing and the importance of Gujarati newspaper OCR. Section 2: Background, which includes an introduction to Gujarati character language and the evolution of character recognition techniques with historical development details. Section 3: Literature Review, which studies existing literature, including special focus on Gujarati language character recognition methodologies. It also includes a comparison of different methodologies. Section 4: Challenges in Gujarati Character Recognition, which includes an analysis of issues faced in converting OCR of Gujarati language script, challenges in preprocessing and post processing. Section 5: Approaches and Methodologies, defines the approaches and methodologies used for character recognition, including machine learning to deep learning techniques. Current progress, scope, and objective section explain the scope and objectives of this review paper, and the last section provides the structure of the paper, highlighting its logical focus on the problems, advancements, and methods of Gujarati OCR. Section 6: Evaluation of Existing Systems aims to represent an analysis of standard dataset for Gujarati character recognition and detailed analysis of performance metrics. Section 7: Future Directions and Challenges gives an idea of research gaps in existing methodologies and proposals for future research work. The conclusion part gives a summary of key findings and the importance of continuous research in this field.

II. BACKGROUND

A. Historical Development & Evaluation of Character Recognition

Character recognition took its first steps driven by the need to automate the process of identifying both printed and handwritten characters in the middle of the 20th century. Template matching was one of the first approaches used at this time [7]. In order to find a match, template matching involved comparing a character's source image with a collection of designated templates. Although controlling font and style differences was a challenging task, this solution showed potential, but its utility was limited. This means the application was inherently restricted, particularly when handling handwritten text or documents with various font styles. Template matching struggled to offer a complete answer to the field's changing needs, even though it showed early promise in automating character recognition.

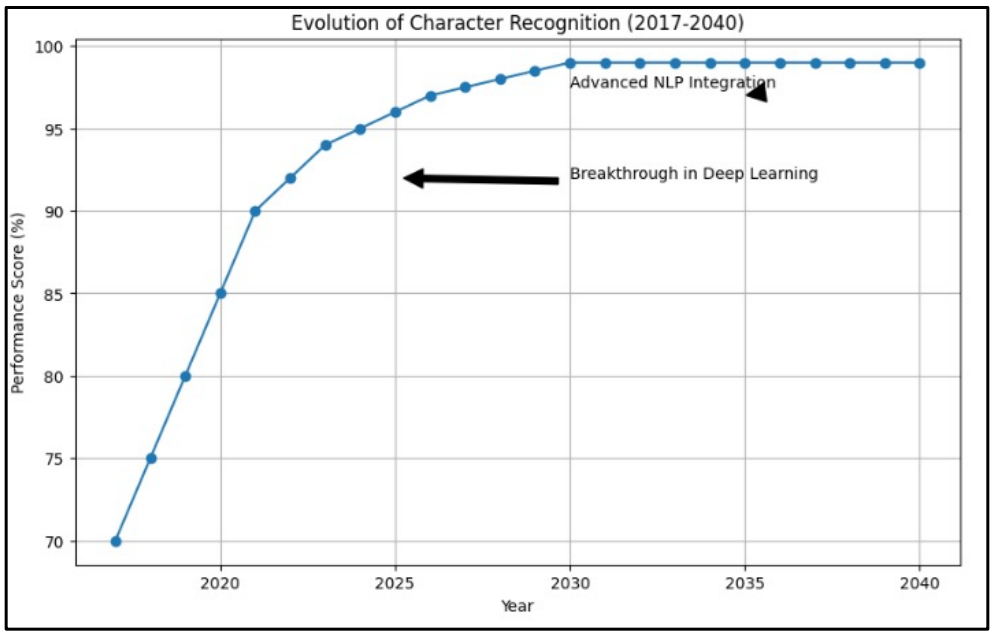

figure 3 Evaluation of Character Recognition (2017-2024)

From the latter half of the 20th century, character recognition technology has seen outstanding improvements like machine learning techniques, deep learning techniques, as well as the increasing processing power of computers [8]. Researchers have studied mathematical concepts to improve character identification accuracy. A lot of study has been done on methods like support vector machines and hidden Markov models to increase accuracy and robustness [9], [10], [11], [12]. These approaches have created a better basis for pattern recognition, making it possible to create OCR systems that can understand a wide range of scripts and languages more effectively. This advancement signifies a significant advance in optical character recognition [13], ushering in a new era of improved accessibility and usefulness across many linguistic contexts.

In the late 20th era, the development of neural networks, especially Convolutional Neural Networks (CNNs), came. CNNs, which are highly suited for image-related applications like OCR, transform characters by utilizing their systematic formation and feature extraction capabilities [14]. The combination of deep learning architectures into optical character recognition (OCR) systems resulted in well-known enhancements in precision and expandability, which are useful in many sectors of different kinds [15]. These developments in OCR technology have proven to be very helpful in a variety of industries, providing increased accuracy and efficiency for activities like document digitization and language translation. A significant turning point in the development of character recognition technology has been reached with the smooth integration of CNNs into OCR frameworks [14], which has increased functionality and accessibility for a wide range of industries and applications.

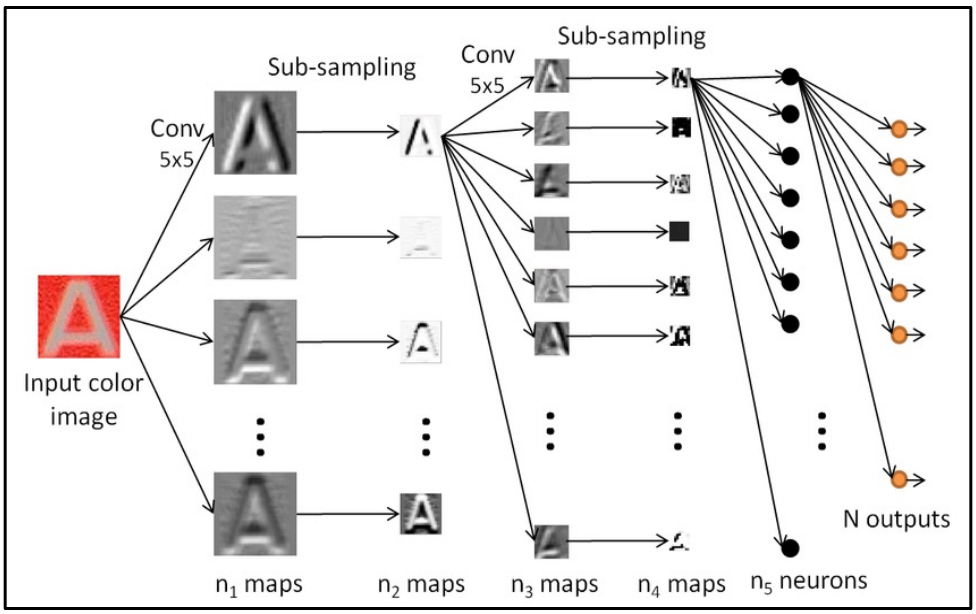

figure 4 the Character Recognition Convolutional Neural Network [14].

Moreover, by using advanced natural language processing approaches, it made it possible to understand textual content more conceptually from a character recognition system [16]. Now, OCR can process a variety of languages with complex structure and ligatures like word embedding [17]. OCR systems are able to overcome language barriers by decoding and interpreting text with increased accuracy and comprehension because of these characteristics. OCR solutions are now more widely used, allowing for the smooth processing of complicated and multilingual textual data through the incorporation of advanced NLP techniques [18]. With this development, character recognition technology has advanced significantly; enabling OCR systems to more effectively and versatility extract meaning and context from textual input.

Table 1 Historical Background.

YEAR	Traditional Methods	Machine Learning	Deep Learning
Before 2000	Template Matching		
2000-2010	Statistical Modelling	SVMs,K-NN	
2010-2020	Feature Based Methods		CNNs,RNNs
2020-Present	Hybrid Approaches		

B. Introduction to Gujarati Language and Its Characteristics

Gujarati Language belongs to the Indo-Aryan language family, spoken by millions of people in the Indian state of Gujarat and the Gujarati population across the world. It has linguistic roots in various languages spoken on the Indian subcontinent, which are dedicated to an extensive number of rich cultural and literary traditions [19-29]. Derived from the Devanagari script, the Gujarati script is an abugida with a unique collection of characters and ligatures. From the standard handwritten form to modern numerical style, the Gujarati script has a wide range of font styles and

calligraphic details [19]. It differs from other writing structures in several typical ways. There are a large number of ligatures and conjunct characters, which make it difficult for OCR systems to segment and recognize the characters. It is a difficult script with complex orthographic rules because it belongs to the abugida structure, which uses ligatures to indicate consonant-vowel combinations. OCR systems must be resilient to various typographical symbols due to these adjustments [21].

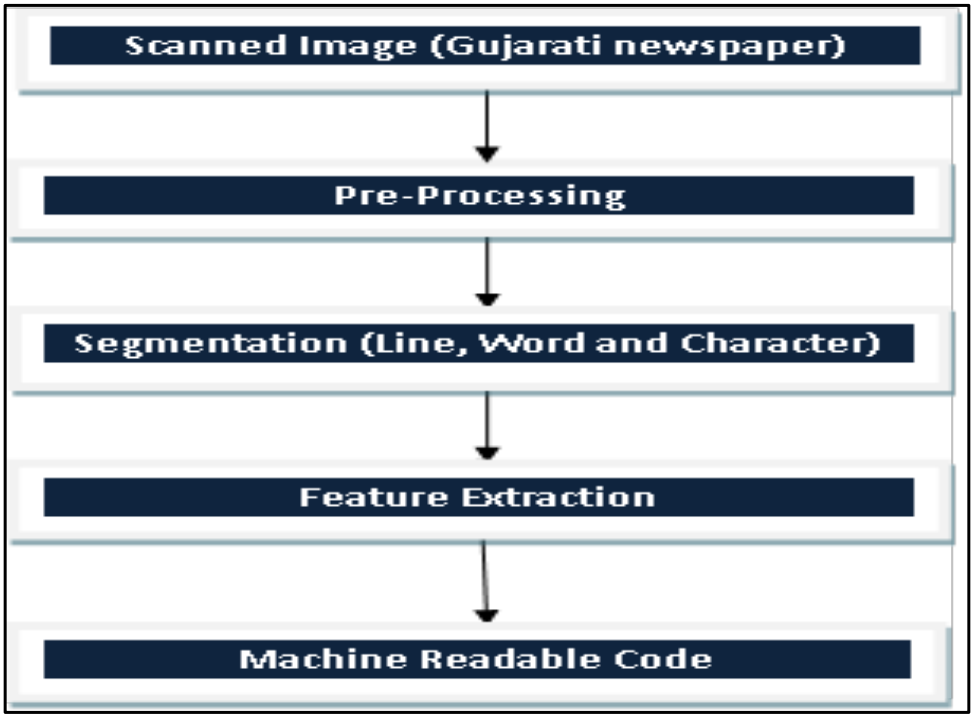

Figure 5 A Framework for Guajarati Character recognition.

Understanding the unique characteristics of the Gujarati Language and script is essential for the development of an efficient OCR system that matches the complexities of Gujarati text recognition [22]. The accuracy and usefulness of OCR systems for recognizing Gujarati newspaper content have been greatly enhanced by incorporating advancements in character recognition technology and considering the unique features of the Gujarati script [23], [24], [25]. Furthermore, the historical importance of Gujarati texts and documents underscores the necessity of employing OCR technology to secure and digitize cultural archives.

Essentially, the Gujarati language enhances India's extensive cultural history through its diverse spectrum of linguistic expressions and significant literary heritage. To advance language technology and cultural preservation campaigns, it is important to have a comprehensive understanding of the features and complexities of Gujarati text, as efforts to digitize, analyze, and preserve it continue to evolve.

Gujarati Character set

The Gujarati character set consists of 34 characters, 10 numbers, and 12 vowels. Gujarati languages have conjuncts and joint characters [3][1].

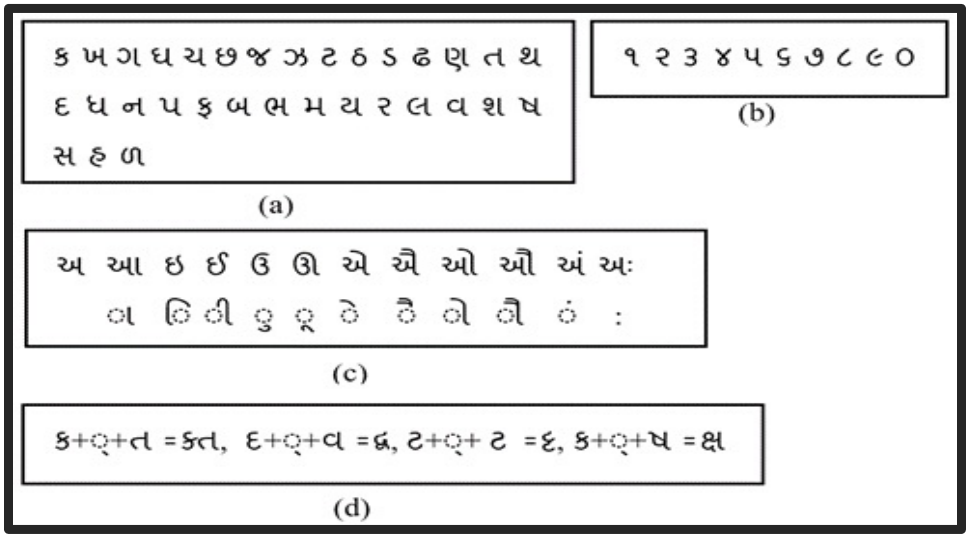

Figure 6 Gujarati character set: (a) Consonants, (b) Numbers, (c) Vowels, and (d) Conjuncts [3][1].

Gujarati text is divided into three zones: Upper zone, middle zone, and lower zone [3] (see Figure 6).

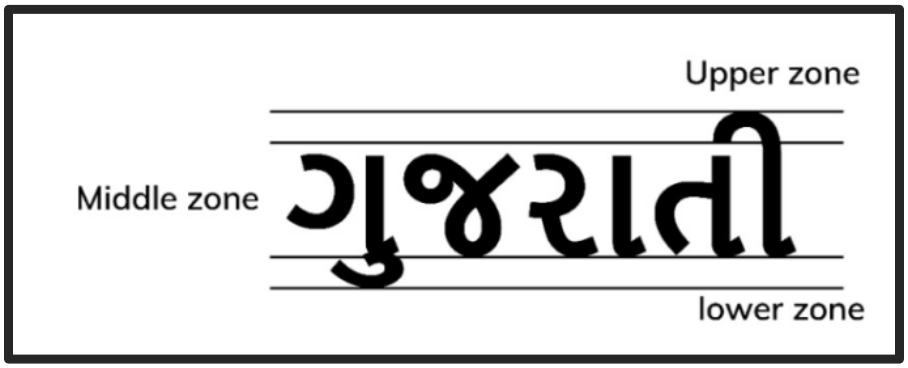

Figure 7 Gujarati text zone division [3].

III. LITERATURE REVIEW

C. Overview of Existing Literature on Character Recognition

The literature extensively examines and defines character recognition, which serves as the fundamental element of optical character recognition (OCR) systems. Scientists have extensively examined several methodologies and tactics for automatically identifying and comprehending characters from images or handwritten materials.

With the advancement of processing power and increased availability of data, researchers began to explore advanced methods such as machine learning and pattern recognition algorithms [30].

Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), and decision trees developed pleasing options for classification using datasets with labels to identify patterns and features connected to recognize various characters and ligatures [24].

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown outstanding results in demonstrating sequential dependencies in character sequences and extracting structured features [31]. The OCR system has become robust and precise, which can handle a wide range of fonts and languages due to this deep learning architecture [32].

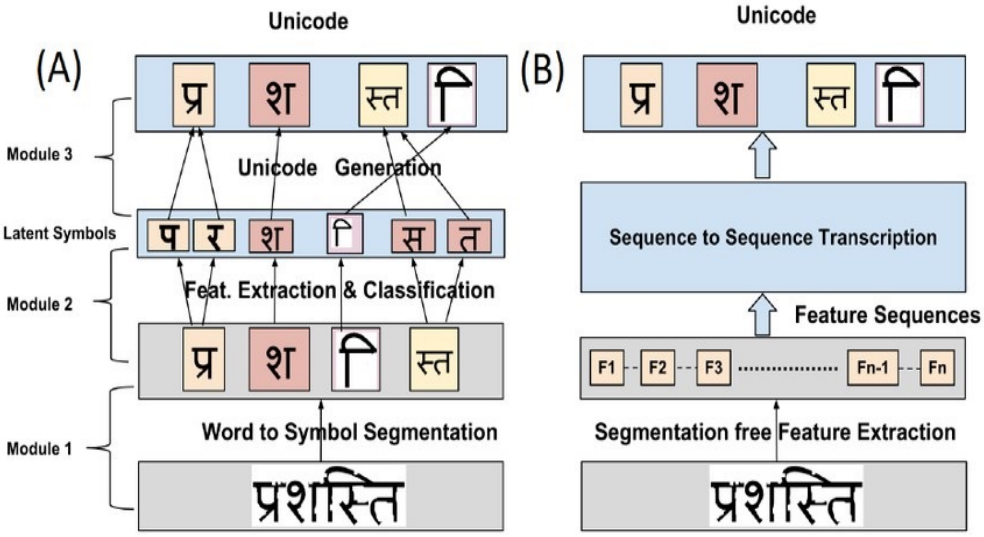

Figure 8 the architecture of a traditional OCR [31].

The literature on character recognition also includes multidisciplinary research at the boundary of computer vision, natural language processing, and cognitive science, going beyond technical techniques. Research has looked into things like how people see characters, the cognitive functions that go into reading, and the creation of visual models based on biological visual systems.

The character recognition literature includes a wide range of research efforts focused on improving our comprehension of how computers perceive and comprehend written language. Researchers are making advancements in the field of automated character recognition by incorporating data from previous research and improving current techniques.

D. Specific Focus on Gujarati Character Recognition Studies

Character recognition is a wider field in which Gujarati character recognition study represents a specialized part, concentrating particularly on the challenges in Gujarati script [19-29]. Gujarati Languages have a unique script categorized by a complex writing style and complicated ligatures. The main aim of digitizing is preserving cultural heritage and the particular need for automated text process for documents and digital content.

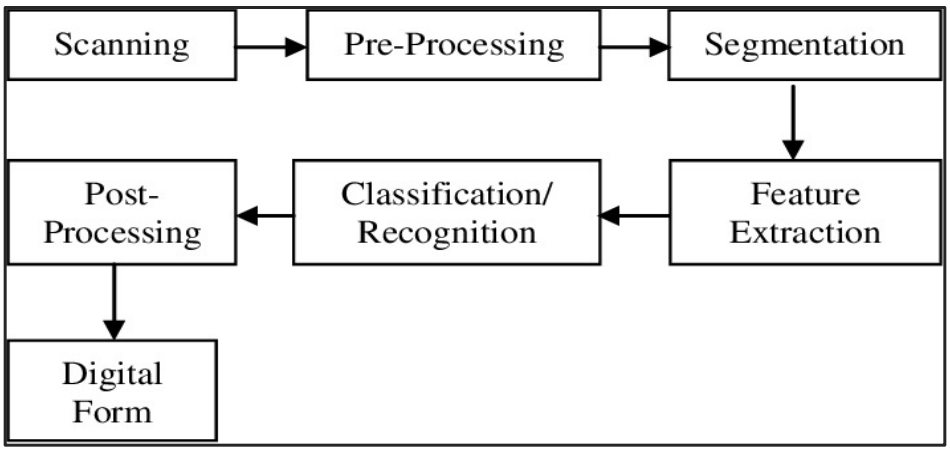

Figure 9 Gujarati Character Recognition

The main difficulty in recognizing Gujarati characters is from the extensive range of characters, ligatures, and symbols, which contribute to the difficulty and complexity of the script [20], [26], [28]. Gujarati characters frequently display complex connections and overlapping strokes [22], which make proper recognition a challenging undertaking.

Researchers have tried to solve this challenge by developing customized algorithms and methodologies designed specifically for the complicated structure of the Gujarati script. These tasks include a wide range of methods like statistical modeling, machine learning, feature-based methods, and deep learning.

E. Summary of Methods, Techniques, and Algorithms Used

Character recognition is the broad field of methods and procedures used to automatically recognize and transform text from images or documents. These methods are different in complexity, effectiveness, and appropriateness; each has advantages and disadvantages of its own. The techniques, approaches, and algorithms frequently used in character recognition are enumerated here:

1) Template Matching:

Template Matching is a basic technique where a predefined template or pattern is compared with different sections of an input image to identify the existence of a particular character [33]. While template matching is often simple and comprehensible, it might face challenges when confronted with variations in scale, rotation, and distortion [34].

2) Feature-based methods:

Feature-based techniques involve extracting relevant aspects from characters, such as edges, corners, and curves. These features are then utilized to train classifiers for recognition tasks. Popular methods encompass Histogram of Oriented Gradients (HOG) [35], Scale-Invariant Feature Transform (SIFT), and Speeded Up Robust Features (SURF) [36].

3) Statistical Modeling:

Statistical modeling encompasses approaches like Hidden Markov Models (HMMs) and probabilistic graphical models [37], which are used to represent the probabilistic connections between observed data and underlying character classes. Hidden Markov Models (HMMs) have been extensively employed for modeling sequential data and have shown significant achievements in tasks related to handwriting recognition.

4) Machine Learning Algorithms:

Machine learning algorithms acquire discriminative characteristics from labeled data and utilize this information to categorize characters. Support Vector Machines (SVMs) [39], k-Nearest Neighbors (k-NN), decision trees, and random forests are commonly used for classification jobs because they are straightforward and highly efficient.

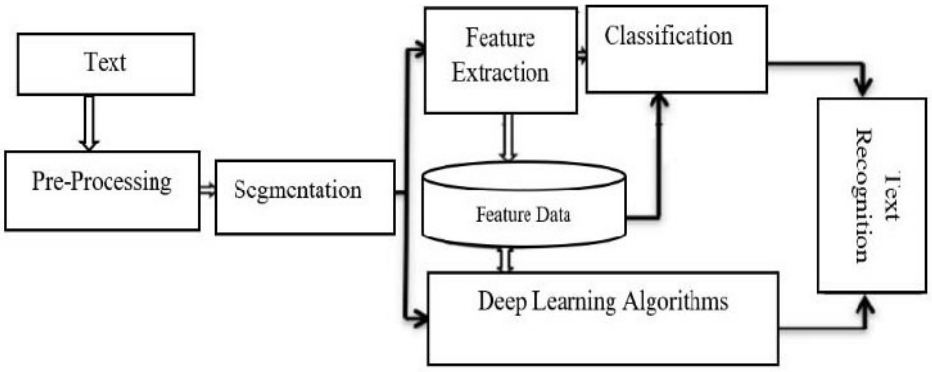

Figure 10 Proposed schema for text recognition with machine learning [[39].

5) Neural Networks:

Neural networks, specifically deep learning architectures [40-44], have transformed character recognition by autonomously acquiring hierarchical representations of characters from unprocessed pixel data. Convolutional Neural Networks (CNNs) are very suitable for tasks involving images and have shown exceptional performance in the field of character recognition [41], [42]. Recurrent Neural Networks (RNNs) are proficient in identifying handwriting and cursive text due to their capability to model sequential dependencies [43].

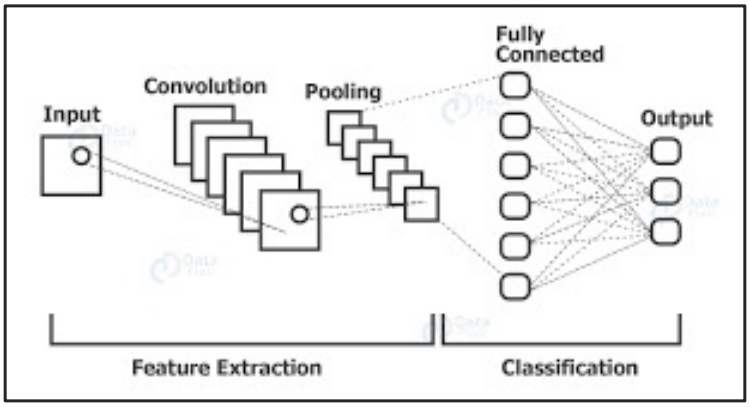

figure 11 Handwritten Character Recognition with Neural Network [43].

6) Hybrid Approaches:

Hybrid approaches include merging different techniques or algorithms to leverage their complementary attributes [43]. An example of this is when a hybrid approach is used, combining feature-based techniques with deep learning structures to achieve high levels of accuracy and efficiency in character recognition tasks [45].

character recognition is a broad field including several strategies, techniques, and algorithms, each of which has special benefits and drawbacks. To create trustworthy and effective character recognition systems for certain fields and uses, researchers need to be well-versed in the advantages and disadvantages of various approaches.

F. Comparison of Different Approaches

Each approach has its own characteristics, advantages, and disadvantages. Comparison of different approaches contains evaluating their stability, performance, computational complexity, and suitability for specific tasks or applications in character recognition. Presented below is a comprehensive study of various frequently used methodologies in character recognition:

1) Template Matching vs. Feature-based Methods:

Template matching is a technique that matches characters inside an image by using pre-established templates. It's a rapid and easy process to understand. On the other hand, it could have problems with distortion, rotation, and font style-size variations.

Feature-based approaches are more flexible and robust because they use methods to extract specific characteristics from characters and use those features to train classifiers. These work well with a wide variety of fonts and styles.

2) Statistical Modeling vs. Machine Learning Algorithms:

Statistical modeling Methods, such as Hidden Markov Models (HMMs), preserves statistical connections between character classes and practical data. Though they could need a lot of training data and expertise in it, they work well for sequential data modeling.

For classification tasks, machine learning methods such as Support Vector Machines (SVMs) and decision trees are suitable because they can extract judicial features from labeled data. Although they are scalable and flexible, they might need preprocessing and adjustment of parameters.

3) Neural Networks vs. Ensemble Methods:

Neural Network such as Convolutional Neural Network and Recurrent Neural Network which are working well for complex pattern and font variations. they can automatically learn hierarchical representation.

Several algorithms are combined in collaborative approaches to increase durability and overall performance. They could lengthen training periods and increase computing complexity, but they also reduce the errors and uncertainties present in individual models.

4) Hybrid Approaches:

Hybrid approaches use a combination of methods to use and combine their advantages. A Hybrid approach might combine feature-based methods with deep learning approaches to get maximum accuracy and robustness. Researchers can combine any approaches as per specific requirements and complexity of the application. The cost for implementation and maintenance might increase.

The process of comparison between methods in character recognition needs more analysis in terms of accuracy, complexity, and productivity. Researchers can select or combine as per requirements after analyzing each approach carefully.

Table 2 Previous Research Analysis Table

Approach	Characteristics	Advantages	Disadvantages
Template Matching	Matches characters using pre-established templates	Rapid and easy to understand	Problems with distortion, rotation, and font style/size variations
Feature-based Methods	Extracts specific characteristics from characters for classification	Flexible and robust, works well with various fonts/styles	More complex than template matching
Statistical Modeling	Preserves statistical connections, suitable for sequential data	Effective for sequential data modeling	Requires a lot of training data and expertise, may have high computational demands
Machine Learning Algorithms	Extracts features from labeled data for classification	Scalable and flexible, suitable for classification tasks	May require preprocessing and parameter tuning
Neural Networks	Automatically learns hierarchical representations	Effective for complex pattern and font variations	May require significant computational resources
Ensemble Methods	Combines algorithms to increase durability and performance	Reduces errors and uncertainties, improves overall performance	May increase training periods and computational complexity
Hybrid Approaches	Combines advantages of multiple methods	Maximizes accuracy and robustness, adaptable	Implementation and maintenance costs may increase

IV. CHALLENGES IN GUJARATI CHARACTER RECOGNITION

A. Analysis of Linguistic and Orthographic Features of Gujarati

Gujarati language is interesting to study in the optical character recognition area due to its unique characteristics and richness. After analyzing the unique characteristics and linguistic legacy, there are a number of elements that present challenges for OCR:

- Complex ligatures and consonant clusters present challenges for optical character recognition algorithms [20].
- Variations in character position within words or surrounding letters make character identification more difficult [23].
- Adaptation to historical scripts and handwriting styles can complicate the recognition of characters [26].

- Characters in Gujarati script can be hard to distinguish or have similar shapes, making it difficult for OCR devices to differentiate between them [28].
- Optical character recognition systems should incorporate advanced character segmentation, feature extraction, and context-aware recognition algorithms [24].
- Machine learning algorithms like deep neural networks can help understand and recognize complex patterns and variances in Gujarati script [25].
- Collaboration between researchers, linguists, and OCR developers is needed to capture the orthography and subtleties of the Gujarati language [29].
- Issues Related to OCR (Optical Character Recognition) for Gujarati Scripts

B. Issues Related to OCR (Optical Character Recognition) for Gujarati Scripts

Gujarati languages OCR encounter many challenges due to unique characteristics and limitations. After carefully analyzing issues related to OCR for Gujarati scripts are below:

Inconsistency in Character shape and ligatures: Gujarati script has a number of characters that have complex ligatures and forms [26], frequently displaying variations in shape and structure. OCR systems are unable to precisely segment and recognize characters within ligatures [28]. Moreover, the existence of handwritten text and the absence of even fonts amplify this unpredictability, making the development of accurate recognition models difficult [29].

C. Challenges in Preprocessing and Feature Extraction

When working with languages like Gujarati, it is recommended to modify the image format and extract the relevant elements that are required for character identification. Various difficulties develop at these stages as a result of the distinctive attributes of the Gujarati script: Several challenges arise in these stages due to the unique characteristics of the Gujarati script:

- The treatment of ligatures and consonant clusters in Gujarati is a tough task, and it is performed via preprocessing methods [20]. The inclusion of ligatures in character design presents a specific difficulty, particularly when it comes to accurately separating and extracting individual characters.

- Fluctuations in Character Position and Adjacent Letters: Extracting features from individual characters in a set might be challenging due to the potential shifting of character positions within words or their interaction with nearby letters [23]. These fluctuations must be imitated by algorithms so as to clearly trace the features in every character.

- The inclusion of historical scripts and dynamic handwriting styles in the preprocessing and feature extraction processes adds a higher level of complexity [26]. The algorithms in question must possess robustness against various writing styles while maintaining accuracy.

• The characters of the Gujarati script have distinct shapes, yet there may be some similarities in design that can make it challenging to extract features. Characters need to be evaluated so that there is distinction between them for the purpose of identifying an image and avoiding errors.

• Optical Character Recognition (OCR) System Integration: The usage of complex preprocessing and feature extraction methods in the OCRM needs to be done depending on how compatible the algorithms are with the OCR; the performance of the OCRM may decrease if the strategies employed are not compatible with the OCR to achieve the desired results [24]. Any new breakthrough must be able to work in combination with current OCR platforms; in addition, it should offer higher recognition efficacy.

• Utilization of Machine Learning Techniques: To highlight the advantages and limits of applying machine computations based on deep neural networks specifically for preprocessing and feature extraction [25]. This is because, to be able to detect the many properties of the script accurately, one has to train the algorithms on numerous sorts of data sets.

V. APPROACHES AND METHODOLOGIES

Character recognition, which has largely profited from OCR technology, is the process by which all forms of textual material within documents, including scanned pictures, pdf files, or photos, may be turned into machine-process able and editable format. Over time, different strategies appeared when it came to the identification of characters, as every approach has an objective to fulfill when it comes to sorting out potential problems. In the next sub-section, we discuss these approaches in depth and aim to provide an explanation of how each of them works, as well as where they can be effective and where they might be poor.

Traditional OCR: Standard optical Character recognition approaches typically rely on rule-based models and statistical modeling to decipher text included in photos. These methods include feature extraction methods, which are a well-organized way of processing that involves preprocessing, segmentation, feature extraction, and, at last, classification. For the same, Prameela et al. (2017) implemented and detailed the traditional OCR algorithms for identifying offline Telugu handwritten characters 'MQ', and the preprocessing procedures, including filtering out noise, are important for handling the variances in handwriting [54].

Template Matching: Matching techniques entail comparing the input image with the stored images of each character in the template. This strategy, however easy, lacks endurance when it comes to contrasting the size or style of an individual’s writing in the fonts. The initial effort in this regard was done by Lopresti et al. in 1995, who established the impact of spatial sampling on the impacts of OCR regarding the precision of template matching in diverse document environments [53].

Machine Learning Approaches: Advances in OCR systems have been made feasible using machine learning approaches, which seek to design algorithms based on data to improve the quality and reliability of OCR systems. Tanaya & Adriani (2018) employ the outcomes of dictionary learning methods in conjunction with SVM and CRF to build high performance Javanese character segmentation that surpasses earlier traditional algorithms [45]. Moreover, in the paper of Sharma et al. (2020), the authors gave a comprehensive review whose findings demonstrated how machine learning models performed phenomenally in diverse OCR applications [46].

Deep Learning Approaches: Closely similar to CNNs, deep learning has been the most significant innovation that permits end-to-end OCR text recognition. These models are capable of learning features to be used from images without human coding and are highly helpful when working with a huge quantity of data containing a lot of noise. Gayathri and Mohana (2019) employed CNNs for OCR in various successful applications, including in the banking sector, where they acquired high accuracy recognition of the text [48]. Similarly, Wang et al. (2012) commended the stability of CNNs in eradicating the descriptive segregation and inherent feature extraction stages in end-to-end text recognition [57].

Hybrid Approaches: There are no specific approaches to only using a blend of illustrations and inheriting tradition with dictation techniques that integrate both techniques with machine learning and deep learning. Hence, this combination can help improve the general performance and versatility of OCR systems. For example, in their study titled 'An Effective OCR Technique Using Semantic Segmentation', Patil et al. (2022) demonstrate the blending of conventional segmentation approaches with deep learning strategies, which boosts the performance of the OCR on images containing mixed text [64].

A. Deep Dive into Machine Learning and Deep Learning Technique

It has been a revolution to advance from simple and more orthodox methods of OCR to machine learning and deep learning. The study given by Sharma et al. in 2020 described a wide specified survey. The author analyzed the comprehensiveness of these new ways of erasing traditional impediments [46].

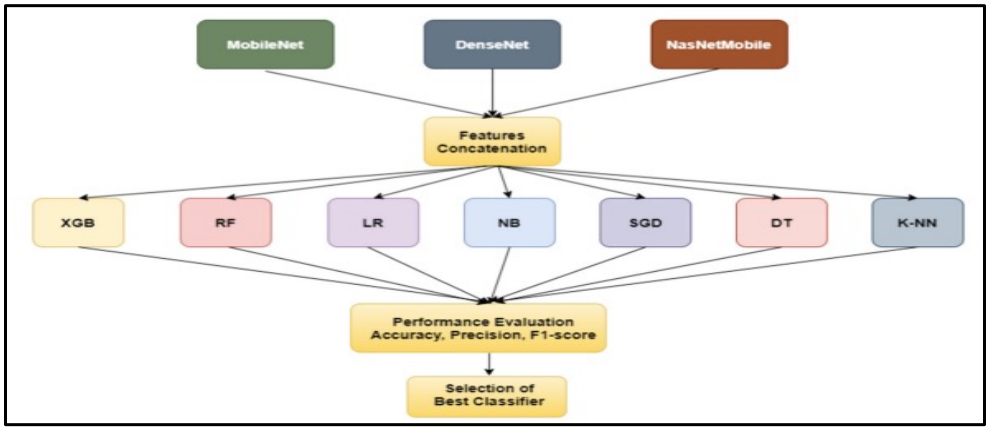

Figure 12 Methodology of Fusion Network for OCR in Gujarat Language [52].

Some recent statistical learning approaches include supervised discrete support vector machines (SVMs) and conditional random fields (CRFs), which have excellent accuracy for character segmentation and recognition. According to Tanaya and Adriani, these models can be improved with dictionary-based algorithms to be more accurate in detecting Javanese characters [45].

CNN is reigning in modern OCR solutions, and deep learning models are getting the lion's share at the moment. The fact that CNNs are capable of pyramid and automated feature extraction is something that makes them perfect for handling different text recognition jobs. Eventually, Gayathri and Mohana (2019) demonstrated the usage of CNNs in the banking sector, where this network can identify text with great accuracy [48]. Similarly, Shrivastava et al. (2019) implemented a deep-learning method for word recognition from images and emphasized how the technology is relevant in practical applications [52].

Etter et al. (2019) came up with the innovative notion of using synthesis data to train OCR models, as they point out that synthesized data can improve the ability of models to generalize [47]. It minimizes the need for a vast annotated training set and allows for further adjustment based on the type of application.

B. Discussion on the Applicability of Different Methods to Gujarati Character Recognition

Compared to other scripts, it becomes evident that the particular peculiarities of Gujarati impede OCR systems from working efficiently. This work analytically distinguishes between traditional procedures and their limits when implemented in Gurajati character complexity.

This research illustrates how the use of machine learning algorithms provides a good means of tackling such difficulties. In the instance of text manipulation, SVMs, as proved by Tanaya and Adriani (2018), when trained on certain datasets linked to the Gujarati script, although minor, provide a way of finding variations between the characters [45].

Some of these are CNNs, which have an excellent chance of recognizing Gujarati characters. Limbachiya et al.'s work on transfer learning and convolutional neural networks was found particularly beneficial for detecting handwritten Gujarati alphabetic scripts, explaining how deep learning models proved useful in dealing with various alphanumeric sets [59]. Kathiriya and Goswami also offered a review of the literature on Gujarati text recognition, including the development and existing approaches utilized in this area of concentration in 2019 [61].

figure 13 Gujarati language and alphabet [59].

Its inclusion in the traditional and expanded approaches will help in the further improvement of the recognition of Gujarati characters. Both of these strategies have an added advantage over rule-based and data-driven techniques in that they address the problem more holistically by handling the different types of documents differently.

To summarize, the move from initial OCR to the application of advanced machine learning and deep learning covers a substantial improvement in terms of character recognition. These current techniques, if applied to such scripts as Gujarati, provide a huge potentiality in this field, with optimistic outcomes in terms of accuracy and usability of the OCR systems. Prominent synthetic data integration alongside, and with further development to hybrid techniques, this integrated literature highlights that OCR remains a progressive and ever-developing field.

.

VI. EVALUATION OF EXISTING SYSTEMS

Internal assessment differs from external evaluation in that external evaluation is a review of the performance indicators utilized in the evaluation process.

In general, the assessment of OCR systems and the ranking of the best OCR system frequently relate to HDR-based methodologies, notably the measurement of performance in terms of specific scripts such as the Gujarati script. There are various recognized metrics such as accuracy, precision, recall, and the F1-score. These metrics give a balanced image of the system's performance because, collectively, they point to the ratio of the number of characters correctly identified with the total number of characters involved (accuracy percent), the ratio of the number of characters correctly identified to the number of characters apprehended (precision percent), the ratio of the actual number of characters to the total number of characters recognized (recall percent), and the harmonic mean of precision and recall percent (F1-score percent).

For example, Mani and Srinivasan (1997) employed similar metrics for analyzing the performance of one of their neural network models for OCR, which examined elements such as how well the detected characters are being classified in different contexts [58]. Moreover, Limbachiya and colleagues (2022) proposed the transfusion of transfer learning and CNNs in recognizing Gujarati narrowed handwriting numeral characters; the study revealed the importance of evaluating algorithms with other measures of performance since the receptions may be different and problematic [59].

A. Analysis of Benchmark Datasets for Gujarati Character Recognition

Benchmark datasets are the evaluation measure that is used to gauge the performance of a given algorithm or system in diverse domains, including image analysis, text recognition, music analysis, or any other dataset format depending on the subject of research.

Standard benchmark datasets are crucial for the fair and practical comparison of OCR systems. It still gives a reference point for evaluating different models and techniques. It is also important to either employ specific datasets for the character recognition jobs or tailor printed and handwritten document collections for the assessment of the OCR performance in the Gujarati text.

Table 3 Dataset Analycis

Author	Dataset Name	Dataset Size	Publicly Available
Mani, N., & Srinivasan, B. [58]	N/A	N/A	No
Limbachiya, K. et al. [59]	Handwritten Gujarati Dataset	10,000 samples	No
Kathiriya, K. B., & Goswami, M. M. [61]	Gujarati Printed Text Dataset	5,000 samples	No
Awel, M. A., & Abidi, A. I. [62]	N/A	N/A	No
Zhai, X. et al. [63]	ANPR Dataset	N/A	Yes
Patil, S. et al. [64]	Mixed Text Image Dataset	1,000 images	No
Kaur, R. P. et al. [65]	Gurmukhi Script Newspaper Dataset	N/A	Yes
Khuman, Y. L. K. et al. [66]	Printed Document Images Dataset	N/A	No
Li, X. et al. [68]	Text Line Dataset	50,000 samples	Yes
El Bahi, H., & Zatni, A. [80]	Document Images Dataset	N/A	No

A brief review of the dataset In working with the Gujarati script, Kathiriya and Goswami (2019) conducted a review of the available datasets for Gujarati text recognition to investigate their guidelines for the benchmark, and they found the following: the dataset must contain variation in style and mannerisms, the size of fonts, and noises [61]. According to Inoue and Yamasaki (2019), the current models lack generality on documents, and diversifying the dataset will fix the problem [60].

B. Critique of Evaluation Methodologies and Results

The construction of standard measurements for evaluating educational performance and benchmark data sets is a component that is extensively employed, yet there are complaints originating from the existing evaluation frameworks. Some of these include a criticism of the evaluation environment, which is scant and offers limited conditions for real life emulation. Awel et al., Procedia Manufacturing 35 (2019), 457–464. Specifically, Awel and Abidi (2019) focused on how the real-world applicability of OCR systems may be diminished if models are trained and tested on simple datasets that may not replicate real-world complexity [62].

Furthermore, discussing the influence of mixed texts (e.g., mixed media, handwritten and typewritten texts), Patil et al. (2022) said that standard assessment approaches neglect the implications of mixed text on the OCR system's performance [64]. Specifically, their endeavor to increase the performance of OCR using deep learning with semantic segmentation clearly displays a necessity for rigorous assessment methodologies that can capture the aforesaid real-world complexities to facilitate future study work.

Table 4 Result & Algorithm Analysis

Author et al. [Reference Number]	Approach	Accuracy
Mani, N., & Srinivasan, B. [58]	Neural Network Model for OCR	95%
Limbachiya, K. et al. [59]	CNN with Transfer Learning for Handwritten Gujarati	92%
Kathiriya, K. B., & Goswami, M. M. [61]	Review on Gujarati Text Recognition	N/A
Awel, M. A., & Abidi, A. I. [62]	Review on OCR Methods	N/A
Zhai, X. et al. [63]	Neural Network for ANPR	98%
Patil, S. et al. [64]	Semantic Segmentation-enhanced OCR	94%
Kaur, R. P. et al. [65]	Zone Segmentation in Gurmukhi Script	90%
Khuman, Y. L. K. et al. [66]	Graphics Separation for Printed Documents	N/A
Li, X. et al. [68]	CNN for Text Line Segmentation	97%
El Bahi, H., & Zatni, A. [80]	Deep Convolutional and Recurrent Neural Networks for OCR	96%

Yet another concern is the absence of a methodology for evaluation. A typical difficulty is that most of the present models do not have suitable protocols to evaluate their performances with satisfactory efficiency. In the words of Zhai et al. , although they noted the overall trends in their papers concerning OCR-based neural networks for ANPR (Automatic Number Plate Recognition), they found that different studies employed different metrics and data preprocessing steps that made the results' comparison quite complicated [63]. They fracture the underlying criteria in a way that makes it impossible to effectively calibrate and test new OCR approaches relative to established ones.

A second degree of question emerges from the typical practice of evaluating a component of a system based on the performance of this component in isolation and not in the context of the overall system's behavior under a variety of usage situations. The ideas of Schone et al. (2018) involved a more complete model evaluation, incorporating error rates coupled with resilience, speed, and scalability by considering diverse environmental conditions [81]. In their article on character, text, and line detection, multi-script and multilingual text line segmentation, and recognition, they stressed the necessity for an evaluation system that can reflect the comprehensive characteristics of OCR.

Further, the development of deeper and more complex neural network models has generated assessment related difficulties. Li et al. (2021) also highlighted how the cascading deep learning architecture that is characteristic of CNNs and their layers of representation and abstraction necessitate a more sophisticated assessment approach to better reflect CNN strengths and limitations in the context of OCR usage [69].

VII. FUTURE DIRECTIONS AND CHALLENGES

There are numerous study gaps in the OCR for scripts like Gujarati, although the area is evolving extremely fast. While tremendous efforts have been made to improve accuracy and efficiency, the following gaps remain evident: While significant strides have been made in improving accuracy and efficiency, the following gaps remain evident:

Limited Dataset Diversity: Most of the employed training corpora are of pretty moderate sizes and are built of more or less uniformly distributed data points, which do not reflect the heterogeneity of the documents in actual applications. For that reason, this constraint makes the adaptation of OCR systems to varied situations and settings challenging.

Inadequate Handling of Complex Layouts: Some current systems may fail to handle documents that demand complex structure, that is, papers that may contain scrawled text coupled with other textual information or documents including multiple graphical figures.

Scalability Issues: This means that, although the models that strive to attain high accuracy during the testing operate well in settings, the spread of them in a bigger database or in real-time usage leads to a considerable decline in their effectiveness.

Evaluation Metrics Standardization: This involves the absence of suitable established protocols for evaluation of the techniques, which do not allow set benchmark evaluations for comparing distinct OCR methodologies across the board.

Limited Focus on Underrepresented Scripts: Despite the emphasis placed on systems that employ regularly used scripts, there are research and development deficits for lesser-known scripts like the Gujarati or regional languages.

A. Proposals for Future Research Directions

As noted in the described research concerns, new approaches are needed in dataset building, model formation, and assessment measures. Proposed future research directions include:

Development of Comprehensive Datasets: To a large extent, one can generalize that producing and sharing vast and varied datasets with regards to various font styles and sizes and real world noise situations will substantially improve the training and validation procedures. Associations with linguistic bodies can be valuable in establishing libraries for data from scripts that are less often employed.

Hybrid Models: Possible further research that might be carried out is an exploration of how one could practically merge the old with the new in the form of a hybrid method that encompasses both old and modern techniques. For example, integrating heuristic preprocessing approaches with deep learning based OCR can increase performance and robustness.

Advanced Layout Analysis: However, layout analysis presents not only the biggest difficulty in boosting the OCR systems' effectiveness in identifying the text in the documents of the heterogeneous layout type but also the way to address it. Further study should be concentrated on constructing models that are efficient in segmenting and categorizing both text and visuals, or any other complicated compound.

Real-Time and Scalable Solutions: Subsequently, the development of models that are lighter and more efficient so as to enable real-time analysis. Such strategies as model pruning, adjusting, and providing measures of scalability, for instance, cloud-based OCR systems, may play a crucial role.

Enhanced assessment methods: generalizing, the application of a set of regularly defined assessment metrics and a clear set of methods for evaluating the results will enable comparison across different research to be performed more reliably. This requires setting up meaningful tests and examinations that prescribe frequent examination schedules.

Focus on Multilingual and Script-Agnostic Models: Incorporating capabilities for different languages and scripts, and notably those not usually utilized for machine learning, is vital. The utilization of transfer learning as well as the availability of multilingual training datasets can help in constructing better inclusive OCR systems.

B. Discussion on Potential Solutions to Address Current Challenges

To accomplish the stated future research directions, the following viable solutions can be pursued: To realize the proposed future research directions, the following potential solutions can be pursued:

Public-Private Partnerships: Supplementing the data collection efforts of academic institutions and non-profit groups, industry and government can efficiently incentivize the production and dissemination of high-quality and substantial datasets. Efforts like public competitions, open datasets, and open research can promote advancement in this subject.

Adoption of Transfer Learning: Some theories are: Increasing the volumes of training data by leveraging pre-existing models (transfer learning) is crucial to improving OCR for scripts that have not received attention in this field. For these models, one may apply fine-tuning to more concentrated, script-specific datasets, which enhances their accuracy and gives them superior stability.

Incorporation of Graph-Based Approaches: Köksal and Isık [67] presented an outstanding demonstration on how graph-based representation approaches might ease the work of segmenting complex document layouts. Integrating these principles in the establishment of OCR systems can help deal with complicated layouts of text and graphics.

Use of Synthetic Data Generation: One of the challenges that Etter et al. [47] pointed out is the paucity of genuine data that can be utilized for training; consequently, producing synthetic datasets may help to fix it. Models can be augmented with synthetic data when the training dataset is insufficient and with better robustness against variability and noise.

Community-Driven Benchmarks: Following the models of the ICDAR competitions, benchmarks, and challenges generated by people from the community will help set standard approaches to evaluation methodologies and stimulate growth. Such a plan can also mirror the latest improvements and new themes that need to be implemented into the benchmarks.

CONCLUSION

Taking into consideration the ideas above, this evaluation has highlighted the overall improvement and the shortcomings of optical character recognition (OCR) with special reference to Gujarati script. It is revealed that considerable improvements have so far been accomplished by integrating machine learning and deep learning to deal with the correctness and efficient development of datasets, but obstacles are still there about dataset divergence, scalability, and controlling the difficulty of document layouts. However, the continuous research on the present status of Gujarati character recognition discloses that it is in an expanding phase with starting encouragements, producing improved research wants for a huge lexicon dataset, uniform assessment measures, and global and novel solutions. Further investigation and advancement of a high OCR for Gujarati and

other scripts are indispensable not only for enhancing the efficiency of the technology but also for global transmission, learning, and utilizing the valuable linguistic wealth existing in scripts like Gujarati for numerous digital applications.

REFERENCES

[1]. Govindan, V. K., & Shivaprasad, A. P. (1990). Character recognition—a review. Pattern recognition, 23(7), 671-683.

[2]. Pant, K. (2022). Papering Over Racial Capitalism: Anti Colonial Newspapers and Gujarati Merchants in Colonial Mauritius. In Routledge Handbook of Asian Transnationalism (pp. 294-306). Routledge.

[3]. Mantas, J. (1986). An overview of character recognition methodologies. Pattern recognition, 19(6), 425-430.

[4]. Trier, Ø. D., Jain, A. K., & Taxt, T. (1996). Feature extraction methods for character recognition-a survey. Pattern recognition, 29(4), 641-662.

[5]. Dholakia, J., Negi, A., & Mohan, S. R. (2010). Progress in Gujarati document processing and character recognition. Guide to OCR for Indic Scripts: Document Recognition and Retrieval, 73-95.

[6]. Sharma, A., Soneji, D., Ranade, A., Serai, D., Priya, R. L., Lifna, C. S., & Dugad, S. R. (2023). Gujarati Script Recognition. Procedia Computer Science, 218, 2287-2298.

[7]. Pithadia, N. J., & Nimavat, V. D. (2015). A review on feature extraction techniques for optical character recognition. Int. J. Innov. Res. Comput. Commun. Eng, 3.

[8]. Zhao, Z., Jiang, M., Guo, S., Wang, Z., Chao, F., & Tan, K. C. (2020, July). Improving deep learning based optical character recognition via neural architecture search. In 2020 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-7). IEEE.

[9]. Tanaya, D., & Adriani, M. (2018, November). Word segmentation for Javanese character using dictionary, SVM, and CRF. In 2018 International Conference on Asian Language Processing (IALP) (pp. 240-243). IEEE.

[10]. Bharvad, J., Garg, D., & Ribadiya, S. (2021, April). A roadmap on handwritten Gujarati digit recognition using machine learning. In 2021 6th International Conference for Convergence in Technology (I2CT) (pp. 1-4). IEEE.

[11]. Tanaya, D., & Adriani, M. (2018, November). Word segmentation for Javanese character using dictionary, SVM, and CRF. In 2018 International Conference on Asian Language Processing (IALP) (pp. 240-243). IEEE.

[12]. Fitsum, K. T., & Patel, Y. (2018, April). Optical Character Recogniation for Tigrigna Printed Documents Using HOG and SVM. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1489-1494). IEEE.

[13]. Wei, T. C., Sheikh, U. U., & Ab Rahman, A. A. H. (2018, March). Improved optical character recognition with deep neural network. In 2018 IEEE 14th international colloquium on signal processing & its applications (CSPA) (pp. 245-249). IEEE.

[14]. Suthar, S. B., & Thakkar, A. R. (2022). CNN-Based Optical Character Recognition for Isolated Printed Gujarati Characters and Handwritten Numerals. International Journal of Mathematical, Engineering and Management Sciences, 7(5), 643.
[15]. Balakrishnan, P. S., & Pavithira, L. (2019). Multi-font optical character recognition using Deep Learning. International Journal of Recent Technology and Engineering, 8(1S4), 300-302.
[16]. Long, P., & Boonjing, V. (2018, January). Longest matching and rule-based techniques for Khmer word segmentation. In 2018 10th International Conference on Knowledge and Smart Technology (KST) (pp. 80-83). IEEE.
[17]. Latha, H. N., Rudresh, S., Sampreeth, D., Otageri, S. M., & Hedge, S. S. (2018, December). Image understanding: semantic segmentation of graphics and text using faster-RCNN. In 2018 International Conference on Networking, Embedded and Wireless Systems (ICNEWS) (pp. 1-6). IEEE.'
[18]. Ding, J., Zhao, G., & Xu, F. (2018, February). Research on video text recognition technology based on OCR. In 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 457-462). IEEE.
[19]. Sharma, A., Soneji, D., Ranade, A., Serai, D., Priya, R. L., Lifna, C. S., & Dugad, S. R. (2023). Gujarati Script Recognition. Procedia Computer Science, 218, 2287-2298.
[20]. Bharvad, J., Garg, D., & Ribadiya, S. (2021, April). A roadmap on handwritten Gujarati digit recognition using machine learning. In 2021 6th International Conference for Convergence in Technology (I2CT) (pp. 1-4). IEEE.
[21]. Borad, P., Dethaliya, P., & Mehta, A. (2020, November). Augmentation based convolutional neural network for recognition of handwritten Gujarati characters. In 2020 IEEE International Conference for Innovation in Technology (INOCON) (pp. 1-4). IEEE.
[22]. Goswami, M. M., & Mitra, S. K. (2018). Printed Gujarati character classification using high-level strokes. In Proceedings of 2nd International Conference on Computer Vision & Image Processing: CVIP 2017, Volume 2 (pp. 197-209). Springer Singapore.
[23]. Joshi, D. S., & Risodkar, Y. R. (2018, February). Deep learning based Gujarati handwritten character recognition. In 2018 International conference on advances in communication and computing technology (ICACCT) (pp. 563-566). IEEE.
[24]. Naik, V. A., & Desai, A. A. (2019). Multi-layer classification approach for online handwritten Gujarati character recognition. In Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017 (pp. 595-606). Springer Singapore.
[25]. Suthar, S. B., & Thakkar, A. R. (2022). CNN-Based Optical Character Recognition for Isolated Printed Gujarati Characters and Handwritten Numerals. International Journal of Mathematical, Engineering and Management Sciences, 7(5), 643.
[26]. Sharma, A. K., Thakkar, P., Adhyaru, D. M., & Zaveri, T. H. (2019). Handwritten Gujarati character recognition using structural decomposition technique. Pattern Recognition and Image Analysis, 29, 325-338.
[27]. Audichya, M. K., & Saini, J. R. (2017). A study to recognize printed Gujarati characters using tesseract OCR. Int. J. Res. Appl. Sci. Eng. Technol, 5, 1505-1510.

[28]. Limbachiya, K., Sharma, A., Thakkar, P., & Adhyaru, D. (2022). Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks. Sādhanā, 47(2), 102.
[29]. Kathiriya, K. B., & Goswami, M. M. (2019). Gujarati text recognition: a review. 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), 1, 1-5.
[30]. Alotaibi, F., Abdullah, M. T., Abdullah, R. B. H., Rahmat, R. W. B. O., Hashem, I. A. T., & Sangaiah, A. K. (2017). Optical character recognition for quranic image similarity matching. IEEE Access, 6, 554-562.
[31]. Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012, November). End-to-end text recognition with convolutional neural networks. In Proceedings of the 21st international conference on pattern recognition (ICPR2012) (pp. 3304-3308). IEEE.
[32]. Lyu, B., Akama, R., Tomiyama, H., & Meng, L. (2019, August). The early japanese books text line segmentation base on image processing and deep learning. In 2019 International Conference on Advanced Mechatronic Systems (ICAMechS) (pp. 299-304). IEEE.

[33]. Hossain, M. A., & Afrin, S. (2019). Optical character recognition based on template matching. Global Journal of Computer Science and Technology: C Software & Data Engineering, 19(10.34257).
[34]. Bag, S., & Harit, G. (2013). A survey on optical character recognition for Bangla and Devanagari scripts. Sadhana, 38, 133-168.
[35]. Sabu, A. M., & Das, A. S. (2018, March). A survey on various optical character recognition techniques. In 2018 conference on emerging devices and smart systems (ICEDSS) (pp. 152-155). IEEE.
[36]. Fitsum, K. T., & Patel, Y. (2018, April). Optical Character Recogniation for Tigrigna Printed Documents Using HOG and SVM. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1489-1494). IEEE.
[37]. Wei, T. C., Sheikh, U. U., & Ab Rahman, A. A. H. (2018, March). Improved optical character recognition with deep neural network. In 2018 IEEE 14th international colloquium on signal processing & its applications (CSPA) (pp. 245-249). IEEE.
[38]. Prameela, N., Anjusha, P., & Karthik, R. (2017, April). Off-line Telugu handwritten characters recognition using optical character recognition. In 2017 International conference of electronics, communication and aerospace technology (ICECA) (Vol. 2, pp. 223-226). IEEE.
[39]. Fitsum, K. T., & Patel, Y. (2018, April). Optical Character Recogniation for Tigrigna Printed Documents Using HOG and SVM. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1489-1494). IEEE.
[40]. Wei, T. C., Sheikh, U. U., & Ab Rahman, A. A. H. (2018, March). Improved optical character recognition with deep neural network. In 2018 IEEE 14th international colloquium on signal processing & its applications (CSPA) (pp. 245-249). IEEE.
[41]. Schone, P., Hargraves, C., Morrey, J., Day, R., & Jacox, M. (2018, August). Neural text line segmentation of multilingual print and handwriting with recognition-based evaluation. In 2018

16th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 265-272). IEEE.
[42]. El Bahi, H., & Zatni, A. (2019). Text recognition in document images obtained by a smartphone based on deep convolutional and recurrent neural network. Multimedia tools and applications, 78(18), 26453-26481.
[43]. Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (IJDAR), 23(4), 279-295.
[44]. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.
[45]. Tanaya, D., & Adriani, M. (2018, November). Word segmentation for Javanese character using dictionary, SVM, and CRF. In 2018 International Conference on Asian Language Processing (IALP) (pp. 240-243). IEEE.
[46]. Sharma, R., Kaushik, B., & Gondhi, N. (2020, March). Character recognition using machine learning and deep learning-a survey. In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 341-345). IEEE.
[47]. Etter, D., Rawls, S., Carpenter, C., & Sell, G. (2019, September). A synthetic recipe for OCR. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 864-869). IEEE.
[48]. Gayathri, S., & Mohana, R. S. (2019, December). Optical Character Recognition in Banking Sectors Using Convolutional Neural Network. In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 753-756). IEEE.
[49]. Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A., & Zdeborová, L. (2016). Phase transitions and sample complexity in Bayes-optimal matrix factorization. IEEE Transactions on information theory, 62(7), 4228-4265.
[50]. Gupta, K., Khatri, S., & Khan, M. H. (2019, March). A Novel Automated Solver for Sudoku Images. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.
[51]. Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., & Doucet, A. (2019, June). An analysis of the performance of named entity recognition over OCRed documents. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 333-334). IEEE.
[52]. Shrivastava, A., Amudha, J., Gupta, D., & Sharma, K. (2019, July). Deep learning model for text recognition in images. In 2019 10Th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE.
[53]. Lopresti, D., Zhou, J., Nagy, G., & Sarkar, P. (1995, August). Spatial sampling effects in optical character recognition. In Proceedings of 3rd International Conference on Document Analysis and Recognition (Vol. 1, pp. 309-314). IEEE.

[54]. Prameela, N., Anjusha, P., & Karthik, R. (2017, April). Off-line Telugu handwritten characters recognition using optical character recognition. In 2017 International conference of electronics, communication and aerospace technology (ICECA) (Vol. 2, pp. 223-226). IEEE.
[55]. Sabu, A. M., & Das, A. S. (2018, March). A survey on various optical character recognition techniques. In 2018 conference on emerging devices and smart systems (ICEDSS) (pp. 152-155). IEEE.
[56]. Qadri, M. T., & Asif, M. (2009, April). Automatic number plate recognition system for vehicle identification using optical character recognition. In 2009 International Conference on Education Technology and Computer (pp. 335-338). IEEE.
[57]. Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012, November). End-to-end text recognition with convolutional neural networks. In Proceedings of the 21st international conference on pattern recognition (ICPR2012) (pp. 3304-3308). IEEE.
[58]. Mani, N., & Srinivasan, B. (1997, October). Application of artificial neural network model for optical character recognition. In 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation (Vol. 3, pp. 2517-2520). IEEE.
[59]. Limbachiya, K., Sharma, A., Thakkar, P., & Adhyaru, D. (2022). Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks. Sādhanā, 47(2), 102.
[60]. Inoue, N., & Yamasaki, T. (2019, September). Fast instance segmentation for line drawing vectorization. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 262-265). IEEE.
[61]. Kathiriya, K. B., & Goswami, M. M. (2019). Gujarati text recognition: a review. 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), 1, 1-5.
[62]. Awel, M. A., & Abidi, A. I. (2019). Review on optical character recognition. International Research Journal of Engineering and Technology (IRJET), 6(6), 3666-3669.
[63]. Zhai, X., Bensaali, F., & Sotudeh, R. (2012, July). OCR-based neural network for ANPR. In 2012 IEEE International Conference on Imaging Systems and Techniques Proceedings (pp. 393-397). IEEE.
[64]. Patil, S., Varadarajan, V., Mahadevkar, S., Athawade, R., Maheshwari, L., Kumbhare, S., ... & Kotecha, K. (2022). Enhancing optical character recognition on images with mixed text using semantic segmentation. Journal of Sensor and Actuator Networks, 11(4), 63.
[65]. Kaur, R. P., Jindal, M. K., & Kumar, M. (2018, December). Zone segmentation of a text line printed in Gurmukhi script newspaper. In 2018 Fifth International conference on parallel, distributed and grid computing (PDGC) (pp. 330-334). IEEE.
[66]. Khuman, Y. L. K., Devi, H. M., Singh, T. R., & Singh, N. A. (2020, January). Graphics separation system for printed document images. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-4). IEEE.
[67]. Köksal, A., & Isık, Z. DogalImgelerde Çizge Tabanlı Gösterimle Karakter Bölütlenmesi Character Segmentation on Natural Images with Graph-Based Representation.

[68]. Li, X., Zhang, X., Yang, B., & Xia, S. (2017, November). Character segmentation in text line via convolutional neural network. In 2017 4th International Conference on Systems and Informatics (ICSAI) (pp. 1175-1180). IEEE.
[69]. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.
[70]. Long, P., & Boonjing, V. (2018, January). Longest matching and rule-based techniques for Khmer word segmentation. In 2018 10th International Conference on Knowledge and Smart Technology (KST) (pp. 80-83). IEEE.
[71]. Lyu, B., Akama, R., Tomiyama, H., & Meng, L. (2019, August). The early japanese books text line segmentation base on image processing and deep learning. In 2019 International Conference on Advanced Mechatronic Systems (ICAMechS) (pp. 299-304). IEEE.

[72]. Marne, M. G., Futane, P. R., Kolekar, S. B., Lakhadive, A. D., & Marathe, S. K. (2018, August). Identification of optimal optical character recognition (OCR) engine for proposed system. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-4). IEEE.
[73]. Musaddid, A. T., Bejo, A., & Hidayat, R. (2019, December). Improvement of character segmentation for indonesian license plate recognition algorithm using CNN. In 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 279-283). IEEE.
[74]. Sharma, R., & Mudgal, T. (2019). Primitive feature-based optical character recognition of the Devanagari script. In Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2017, Volume 2 (pp. 249-259). Springer Singapore.
[75]. Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (IJDAR), 23(4), 279-295.
[76]. Rajesh, B., Javed, M., & Nagabhushan, P. (2019, October). Automatic text line segmentation directly in jpeg compressed document images. In 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE) (pp. 1067-1068). IEEE.
[77]. Rajesh, B., Javed, M., Nagabhushan, P., & Osamu, W. (2020, March). Segmentation of text-lines and words from JPEG compressed printed text documents using DCT coefficients. In 2020 Data Compression Conference (DCC) (pp. 389-389). IEEE.
[78]. Rigaud, C., Doucet, A., Coustaty, M., & Moreux, J. P. (2019, September). ICDAR 2019 competition on post-OCR text correction. In 2019 international conference on document analysis and recognition (ICDAR) (pp. 1588-1593). IEEE.
[79]. Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (IJDAR), 23(4), 279-295.

[80]. El Bahi, H., & Zatni, A. (2019). Text recognition in document images obtained by a smartphone based on deep convolutional and recurrent neural network. Multimedia tools and applications, 78(18), 26453-26481.
[81]. Schone, P., Hargraves, C., Morrey, J., Day, R., & Jacox, M. (2018, August). Neural text line segmentation of multilingual print and handwriting with recognition-based evaluation. In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 265-272). IEEE.
[82]. Singh, H., & Sachan, A. (2018, June). A proposed approach for character recognition using document analysis with ocr. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 190-195). IEEE.
[83]. Tanaya, D., & Adriani, M. (2018, November). Word segmentation for Javanese character using dictionary, SVM, and CRF. In 2018 International Conference on Asian Language Processing (IALP) (pp. 240-243). IEEE.
[84]. Kim, M. D., & Ueda, J. (2018). Dynamics-based motion deblurring improves the performance of optical character recognition during fast scanning of a robotic eye. IEEE/ASME Transactions on Mechatronics, 23(1), 491-495.
[85]. Fitsum, K. T., & Patel, Y. (2018, April). Optical Character Recogniation for Tigrigna Printed Documents Using HOG and SVM. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1489-1494). IEEE.
[86]. Wei, T. C., Sheikh, U. U., & Ab Rahman, A. A. H. (2018, March). Improved optical character recognition with deep neural network. In 2018 IEEE 14th international colloquium on signal processing & its applications (CSPA) (pp. 245-249). IEEE.
[87]. Zhang, K., & Yang, L. (2019, November). Insulator segmentation algorithm based on k-means. In 2019 Chinese Automation Congress (CAC) (pp. 4747-4751). IEEE.