International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

A Study of Techniques and Challenges in Text Recognition Systems

Gurvir Kaur[1], Ajit Kumar[2]
[1]Department of Computer Science,
Punjabi university,
Patiala, India
[2]Department of Computer Science
Multani Mal Modi College
Patiala, India

Abstract—The core system for Natural Language Processing (NLP) and digitalization is Text Recognition. These systems are critical in bridging the gaps in digitization produced by non-editable documents, as well as contributing to finance, health care, machine translation, digital libraries, and a variety of other fields. In addition, as a result of the pandemic, the amount of digital information in the education sector has increased, necessitating the deployment of text recognition systems to deal with it. Text Recognition systems worked on three different categories of text: (a) Machine Printed, (b) Offline Handwritten, and (c) Online Handwritten Texts. The major goal of this research is to examine the process of typewritten text recognition systems. The availability of historical documents and other traditional materials in many types of texts is another major challenge for convergence. Despite the fact that this research examines a variety of languages, the Gurmukhi language receives the most focus. This paper shows an analysis of all prior text recognition algorithms for the Gurmukhi language. In addition, work on degraded texts in various languages is evaluated based on accuracy and F-measure.

Keywords- Text Recognition System, OCR, NLP, Typewritten Text, Gurmukhi language.

I. INTRODUCTION

The use of digital information is increasing every day, and it has become an integral part of everyone's life. OCR is a part of Natural Language Processing (NLP) that helps in digitalization by scanning paper documents. The characters in the documents are recognised in a variety of methods, and the records are then turned into editable or searchable formats. Computers all throughout the globe deal with a vast quantity of paper-based documents in the form of images. [1]. Segmentation, feature extraction, and classification are only a few of the essential processes involved in OCR text recognition [2]. The basic goal of OCR is to convert printed and online/offline handwritten text images into a machine-editable format that is as human-readable as possible. Its main focus is on alphanumeric or other character recognition that has been optically processed [3]. OCR encompasses artificial intelligence, pattern recognition, digital image processing, database systems, and human-machine interfaces [4]. Scanned newspaper clippings, books, and magazine pages can be used to make document images [5]. As a consequence, OCR technology turns a document's picture into ASCII/UNICODE code. Among the most popular OCR applications include automatic identification of official handwritten/printed documents, automatic recognition of number plates, automated filling forms and a broad range of other commercial applications. In general, both offline and online, two methods for detecting handwritten/machine-printed languages in character recognition. Images of pre-written text that have been scanned in paper are known as offline documents [6] [7]. Typing on a keyboard or using an electronic pen on a computer captures data in online records. Here's a closer look at the many types: Machine Printed Text Recognition System: The Machine Printed Text Recognition System is the most well-known, since it translates machine-printed text from physical paper to an electronic version. For example, data produced on typewriters or data developed on computers and printed on printers. Artificial intelligence might be used by these gadgets to decode the text pattern and convert it from physical to digital form. It will boost output while also influencing digitalization [8]. The produced data can be captured by scanning or taking photographs with a camera. Numerous OCR systems have been developed for machine-produced text in a variety of languages [9].

A. Offline Handwritten Text Recognition System Handwritten samples are scanned using scanners or shot with the camera. The fundamental flaw in this identification approach is handwriting variation, as people's writing styles differ [10] [11]. OCR technologies for handwritten text are also available in a variety of languages.

B. Online Handwritten Text Recognition System Text made with an electronic pen is referred to as handwritten on the internet. In this scenario, the variation in the person's writing is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

a cause for concern. There's also the issue of the pen's pressure on the device, which affects writing style. Most organizations now use online handwritten text [12] [13] recognition systems because they used to collect data/information in the form of online handwritten text like a signature for a successful delivery for some e-commerce sites or some feedback services because they used to collect data/information in the form of online handwritten text like a signature for a successful delivery for some e-commerce sites or some feedback services.

Typewritten text is a type of machine-printed text that is commonly seen in ancient and historical documents. Because of the complexity of typewritten text, only a few techniques have been developed. This section also goes through the specifics of typewritten text recognition systems.

1.1 Typewritten Text Recognition Systems

Typewritten Text recognition software reads typewritten text and converts it to a digital version. This material was written using typewriters, but also share certain characteristics with printers. These systems work with samples that have been photographed or scanned using scanners. In terms of shape and size, most typewritten papers use a comparable typeface. Identifying old typewritten materials with low quality or damaged text, on the other hand, is the most challenging task [14]. The typewritten text is of low quality and variable consistency, and many of these documents have degraded due to the age of the paper and ink used. The typewritten text frequently contains non-uniform characters, some darker or fainter than others, depending on the amount of force necessary to press the typewriter key. Working with typewritten papers is more challenging than working with machine printed materials because text degradation demands pre-processing. Typewriters are fast becoming outdated during this time period. Historical data and some legal and sensitive information from the past are still only available on paper and were typewritten. It is necessary to develop a text recognition system that can recognize typewritten letters and alter them as needed. It's a subclass of machine-printed text, and some of its uses are shown in Figure 1.

The application of typewritten/machine printed text can be in the field of banking, health care, legal, etc. are as follows. The banking business is the most significant and requires special attention, and OCR assists in this field. It's widely used to convert checks into digital text without the need for human intervention. It will help in the process's speeding up and automation, decreasing processing costs. They also use receipt imaging to keep track of the organization's financial records, which includes a record of all transactions and payments, as well as some autonomous and government entities. Organizations use OCR to collect data and information and to assess it. Other industries with significant volumes of printed material and a requirement for a speedy identification system include legal, health care, and document database systems. ATMA is also a software that is utilised by local visitors and travelers. This application captures native language book pages, banners, signboards, and other images, then uses an OCR to extract text from them. It is also possible to convert the recovered text into the traveler's native language. Finally, digital libraries are major applications in which OCR may aid in a variety of ways.

1.2 Basic Process of Typewritten Text Recognition System

The typewritten Text Recognition system examines images acquired with cameras or scanned using scanners. The typewritten text frequently contains non-uniform characters, some darker or fainter than others, depending on the amount of force necessary to press the typewriter key. The text recognition system for typewritten, machine-printed, or other recognition systems worked in a similar way to the handwritten, machine-printed, or other recognition systems. The collection of samples, which might be scanned documents or images taken with cameras, is the first step in the identification method. The stages for further processing include pre-processing, segmentation, feature extraction, and classification, as shown in fig 2. Post-processing is occasionally performed if the recognition system demands it. Below is a detailed description of each level:

- *Sample Acquisition:* The first stage in the recognition system is to gather samples, which may be done using scanners or cameras, and then save them in formats such as.bmp,.jpeg, and.tiff. The source image might be grayscale, RGB, or binary.

- *Pre-processing:* Pre-processing is required to improve the image's quality before it can be utilised for further analysis. The goal of pre-processing is to guarantee that all critical document layout analysis and classification activities are completed appropriately [15]. The initial step in this process is to eliminate any undesirable noises or other modifications that have happened as a result of the collection and subsequent deterioration of the samples. The OCR system's outstanding accuracy is achieved by the use of binarization, noise reduction, normalisation, thresholding, baseline detection, and other pre-processing techniques. Many difficulties, such as blurring, external noise, missing line segments, and distortions at corners, are introduced in these example images. Filtering, morphological procedures, and noise modelling methodologies, as well as erosion and dilation, can help minimize all of these hurdles [16].

- *Segmentation:* The next crucial step is segmentation, which entails breaking down documents into lines, characters, or words in order to improve recognition rates. It mostly applies to text and decomposes the text from visuals, pictures, or figures. It deals with two types of segmentation: external and internal. In

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

general, any complex text picture or language may be divided into text and non-text parts using external/holistic segmentation [17]. External segmentation is important in page layout analysis, also known as document analysis, separated into the structural and functional analysis. External segmentation divides the structural and functional analysis, which is also a vital stage, into page layout analysis, also known as document analysis. Internal segmentation, on the other hand, breaks down the image's words into individual letters or symbols as a group of pixels with some more significant associated properties.

- *Feature Extraction:* The purpose of feature extraction is to give more relevant and matching qualities that allow a letter or symbol to be recognised. The input image's predictive capability is improved by using a combination of feature development and feature selection. Feature extraction challenges are tackled by constructing a data analysis approach and selecting discriminative characteristics to simplify problematic data sets [18] [19]. There are three types of characteristics/features that are widely used. Characteristics in terms of statistics, structure, and global transformations. Topological and geometrical properties of an image, such as dots, loops, and endpoints, as well as statistical variables generated from the picture region using numerical measurements such as zoning, projections, histograms, crossings, and distances, are computed using structural characteristics. Using global transformations, various transformation methods such as Fourier transforms, global transforms, wavelets, and moments may be utilized to characterize the visual signal in a compact manner.

- *Classification:* The classifier or recognizer detects the isolated symbols/characters once features are retrieved, which increases the OCR system's performance. Before choosing and assigning likely output labels, the classifier is used to categories the input features. The primary classification methods include the hidden Markov model, nearest neighbour classifier, support vector machine, decision tree classifier, and others. In general, the paper portrays differentiates between two approaches: holistic/top-down and analytical/bottom-up [16]. The comprehensive approach recognises the complete word and eliminates segmentation problems. The analytical process, on the other hand, begins at the character level and develops to the generation of more understandable language. The system's powerful algorithms, on the other hand, produce segmentation issues. The post-processing phase is used to correct some errors in the classifier.

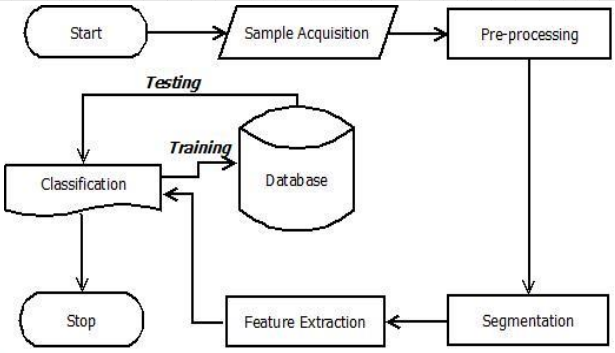


Figure 1. General Framework for Typewritten Text Recognition System

Post-processing: Using semantics and domain knowledge, the dictionary, and other tools, post-processing is used to improve recognition results. It's the last step, and it displays the recognized text in a structured format.

This template provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. PLEASE DO NOT RE-ADJUST THESE MARGINS. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. RELATED WORK

Wherever The following are some of the available approaches for recognizing typewritten documents in various languages.

Pletschacher et al. [20] described a semi-supervised clustering approach for detecting severely deteriorated ancient typewritten texts. Abubacker and Gandhi [21] proposed a more sophisticated way to dealing with problems about character breakage in text documents. The focus of this study is on Tamil typewritten text. Based on an identification and recognition system, Cao et al. [22] proposed the Hidden Markov Model (HMM) for handwritten and typewritten text extraction from images. In this work, type-independent parameters were collected using the Arabic language For machine typed Text recognition, Vukovi and Arizanovi [23] proposed a segmentation algorithm. Because it involves some threshold value alterations and is suited to the English language, this approach is not totally automated. They compared line, word, and character segmentation as well as three different

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

segmentation scenarios. Mahmood and Srivastava [24] proposed a new segmentation method that includes three main modules: connecting component detection, connected component edge detection, and segmentation. For testing purposes, a dataset of 115 text graphics was gathered from a variety of books, newspapers, and journals. This method's performance was assessed using two different performance metrics: precision and recall, which gave 87.36 percent and 84.75 percent, respectively. The additional study on the Urdu language was done by [25] [26] [27].

Panda and Tripathy [28] proposed a novel approach for the Indian language Odia, based on Unicode mapping and template matching. This study employed binarization, skew detection, and correction to pre-process scanned input data. They used the threshold technique for binarization, using a threshold value of 128 as the value. To avoid mistakes, skew detection and correction techniques were used, as well as the elimination of boundaries. They evaluated the proposed method on paper pictures, scanned images, and camera images, and found it to be 100% accurate, 97.07% accurate, and 91.17% accurate, respectively. This suggested approach achieved an overall accuracy of 97.87%. AI techniques are used in other Indian languages such as Devnagri [29][29][30] and Tamil [31].

Khorsheed [32] used a segmentation-free strategy to provide a new feature extraction method for recognising cursive typewritten text. When testing was done in this study, the feature vector of seven dimensions was retrieved, and it was then assessed that this suggested method performed considerably better than his previously created [33] technique and had good recognition results. Retsinas et al. [34] demonstrated an effective recognition system with little user participation. This method has a substantially lower error rate, and its performance is excellent.

According to the literature discussed in this paper, the English language accounts for 45% of all effort done on typewritten text recognition systems. Arabic, Tamil, Urdu, and Odia, for example, are regarded to a lesser extent, with 22% for Arabic and 11% for the rest of the languages mentioned in Figure 3. This analysis, on the other hand, is not limited to the quantity offered and is solely based on the work presented. This study clearly demonstrates that typewritten material is not being taken into consideration as it should be.

Researchers concentrated their efforts on machine-printed and handwritten text for the most part. India is a big country with many different languages spoken in various regions. The Punjabi language, for example, is spoken in northern India (especially in Punjab) and written in the Gurmukhi script (explained in the following section). When analyzing typewritten text, this research also considers the Gurmukhi script and explores its qualities, relevance, and problems.

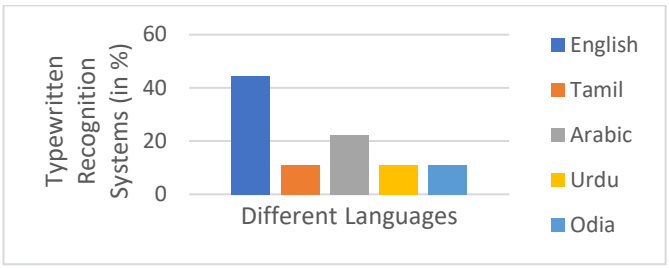


Figure 2. Typewritten Text Recognition Systems for Different Languages.

1. Gurmukhi Script

The Punjabi language, the world's 12th most frequently spoken language, is written in the Gurmukhi script. It's made up of 35 consonants from Punjabi-painted alphabets, comprising three vowel holders and 32 consonants, and it's written in horizontal lines from left to the right. A phonetic sound is represented by each letter. The Gurmukhi script has a different alphabetical order than the English alphabet. Gurmukhi Akhar is organized in five horizontal and seven vertical rows with particular phonetic qualities and is based on groups with certain commonalities. Depending on its horizontal and vertical location, each letter has a unique set of properties. Some letters are spoken with the tongue curved back to touch just behind the ridge on the roof of the mouth or to touch the back of the upper teeth. Letters can either be spoken with a puff of air or by holding back air. Some characters have a nasal tone to their voices.

A horizontal line connects the characters in the words of Gurmukhi Script at the upper side, which is called headline, and blank spaces separate words. Generally, a word in the Gurmukhi script is divided into three different regions (i) Upper Zone, (ii) Middle Zone, and Lower Zone, as shown in Fig 4. The upper zone denotes the region above the headline, where vowels reside, while the middle zone represents the area below the headline where the consonants and some sub-parts of vowels are present. The lower zone represents the area below the middle zone where some vowels and certain half characters lie in consonants' feet. All the characters in the middle zone touch the headline at least once. If present in character, the vertical line is present mainly on the rightmost end of the character.

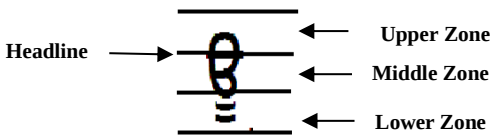


Figure 3. Different Zones of Gurmukhi Character.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

2.1 Problem Statement

In the digital age, any type of material should be available online for a variety of reasons. As a result, OCR has become a must in today's environment. Previously, typewriters were used by the majority of companies and educational organizations. Typewriters are used by several government agencies to type documents on paper. To save time and convert typewritten data into an editable format that fits certain requirements, an OCR system must be developed. OCR has become a well-known research platform due to the range of languages, typefaces, and styles of text data available in documents. Other recognition algorithms have been developed, and they have done well in Indian languages, but typewritten text still requires OCR. In northern India, notably in Punjab, the Punjabi language is written in Gurmukhi Script. Existing OCRs for Gurmukhi text recognition have a number of drawbacks, including (a) only identifying high-quality samples and (b) scanning inputs with some restrictions. Because the majority of government agencies in Punjab typewritten documents in Gurmukhi Script, an automated recognition system that recognizes typewritten Gurmukhi text and converts records to digital format for easy processing is required. The main goal of this project is to create a system that can recognize typewritten Gurmukhi text, recognize it accurately like a person, and extract information from it.

Jindal et al. [35] looked at Gurmukhi text for touching characters. Mangla and Kaur [36] presented a method for segmenting handwritten Punjabi text into touching and breaking letters. Jindal et al. [37] presented a structural features-based technique for recognizing Gurmukhi text that is damaged. [38] [39] [40] [41] suggested another text recognition method for the Gurmukhi language. For the Gurmukhi language, the majority of the effort is done on machine-printed Text, whereas Gurmukhi Text for typewritten texts receives far less attention. As a result, there is a need to create a recognition system for Gurmukhi typewritten texts.

2.2 Challenges in Gurmukhi Typewritten Documents

Some of the challenges found in the sample dataset are given below:

- *Underlines in some text*

Because this text was typed, part of the information has undesired underlines. This underlining will degrade the segmentation efficacy of the OCR/Recognition system, resulting in poor performance.

- *Contains English and Numeric Content along with Gurmukhi*

Because the data comes from a thesis written in the Gurmukhi language, it does not contain Gurmukhi text exclusively, and there is a mix of numeric and Gurmukhi in certain places (for example, see fig 5(a)). There is also some material in 'English Language' in this data. As a result, determining the text's language may be this project's most difficult challenging task.

- *Heavily Printed Characters*

As illustrated in fig 5, typewritten Gurmukhi text likewise faces the difficulty of highly printed text (b). It's difficult to extract a characteristic from a highly printed character. It's a blob of pixels the same height and width as the original character, with no ascenders or decanters to help identify them. Generally, heavily printed characters also touch with neighboring characters, i.e., even falling in the touching character category.

- *Broken Characters and Light Printed Characters*

A single character is split into many components in this type of degraded writing [42]. It's also been shown that fragmented characters are more likely to produce mistakes than touching or highly written ones. This might be a natural result of the fact that there are more white pixels than black pixels on the page, even in text regions. As a result, turning a black pixel into a white pixel loses more information than converting a white pixel to a black pixel. Excessive fragmentation can ruin a whole sentence, making it difficult for humans to recognize. Only a few pixels of a character remain in difficult situations, not enough for a human to identify the character in isolation. Due to the presence of broken characters, as shown in fig 5 (c), any OCR's performance may further decrease.

- *Touching characters*

This is by far the most prevalent flaw in typewritten material. The most important aspect of detecting touching characters is accurately segmenting them, determining where the touching pair of characters must be divided. Every OCR must perform admirably when it comes to the crucial duty of separating them. The accuracy of an OCR is highly reliant on the segmentation process' precision. The presence of touching characters in any document substantially reduces OCR recognition accuracy.

- *Unusual space between words*

It also creates segmentation problems.

- *Splashed Ink/Characters*

Ink splashes on the paper occasionally, resulting in illegible letters, and it also creates markings on the page that resemble line punctuation marks. The printed lines become harder to comprehend.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

- *Skewness and Headlines*

Two or more letters/symbols may overlap if characters or words are slanted to the left or right. Characters are always related to surrounding characters in Gurmukhi Script owing to the existence of the headline; however certain characters do not have headlines. Some characters appear to be the characters without headlines due to broken wording. So, the identification of these characters is difficult.

2.3 Literature Review

This section describes the current Gurmukhi text recognition systems. The amount of work done on the Gurmukhi language, according to the findings of this study, is exceedingly little and deserves experts' attention. Another big issue is the many types of deterioration. The following is a list of some of the work that has been completed in relation to the same:

Jindal et al. [35] looked at touching characters in Gurmukhi text and classified them into three groups based on their structural features and the zones to which they belonged: (a) middle zone, (b) Upper Zone, and (c) Lower Zone. This algorithm recognized and segmented 92-95% of touching characters in the middle zone. The algorithms proposed in this paper applied only to machine-printed text. Madan et al. [38] presented a generic feature extraction technique for Gurmukhi and other Indic languages. The refinement of this algorithm is required to handle variations in fonts and their sizes. Jindal et al. [39] discussed the different kinds of degradations present in the printed documents, so in this work, they have identified various degradations present in the Gurmukhi Documents. These degradations were broken characters, touching characters, and heavily printed characters. The authors also discussed some solutions to these degradations, but there is still a work scope for Gurmukhi and other languages for this kind of solution.

Jindal et al. [37] presented a structural features-based technique for recognizing Gurmukhi text that is damaged. They computed structural elements such as sidebar presence, half-sidebar presence, headline, number of junctions with headlines and baseline, Aspect Ratio, Profile Direction Codes, Directional Distance Distribution, and Transition Features. They then segregated data using their projection-based technique [41]. They analyzed scanned documents from books, newspapers, periodicals, and other sources. Single characters were acquired after segmentation was performed to the data to segmentation touching characters. The performance of the KNN and SVM classifiers was then evaluated using different combinations of features utilizing both classifiers. SVM obtained 91 percent accuracy, which was the best accuracy reached using multiple combinations, while KNN achieved 83.6 percent accuracy, the highest accuracy achieved using KNN.

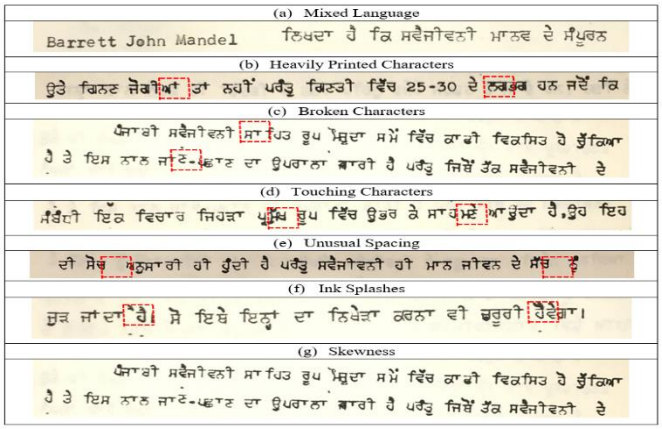


Figure 4: Challenges in Typewritten Gurmukhi Scripts

The Gurmukhi OCR system was provided in a novel way by Lehal [40]. Because this method was not reliant on font or toughness, it performed well on older manuscripts. In this recognition system, binary tree, KNN, and SVM classifiers were used in parallel and serial modes. Furthermore, the corpus-based Weighted Voting Method was utilized to aggregate the outputs of similar classifiers. These classifiers were merged to compensate for the shortcomings of the individual classifiers while maintaining their strengths, resulting in improved performance. Using the structural feature-based approach, this work also addressed issues such as broken characters. The author tested this system on a dataset of 31 pages, which contains 42650 characters with 10 different fonts. This system's recognition accuracy was 98%, which was the best compared with other classifiers.

Mangla and Kaur [36] presented a method for segmenting handwritten Punjabi text into touching and breaking letters. The authors of this paper scanned handwritten text documents, then used an intensity-based threshold technique to transform scanned pictures into binary images. The algorithm's second stage was to remove the header file, followed by character segmentation based on vertical profile projection. To begin, a header line was removed, and the frequency of black pixels in each row was determined using the Horizontal Profile Projection method. Then the row with the black pixels was chosen and used as a header, and all of the 1s were changed with 0. Before segmentation, they utilized the adjacent pixel method to locate broken characters. In this study, structural characteristics were employed to identify touching characters, followed by segmentation. For both touched and broken characters, this approach obtained a 95 percent accuracy rate.

Jindal et al. [41] developed a technique for segmenting touching characters in printed Gurmukhi script that considered all three zones. Multiple horizontal overlapping lines were utilized in this piece to achieve this. Fifty-four scanned documents from various printed newspaper articles were used to segment horizontally overlapping lines, with 95 percent accuracy in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

TABLE 1: DIFFERENT GURMUKHI TEXT RECOGNITION SYSTEMS

Author	*Dataset Source*	*Dataset Size*	*Dataset Type*	*Quality of Sample*	*Device used*	*Degradations*
Jindal et al. [35]	various books and magazines	500 documents	Machine Printed	300dpi	Scanner	touching characters
Jindal et al. [39]	Not Given	Not Given	Typewritten	200dpi	Camera	broken, touching, and heavily printed characters.
Jindal et al. [37]	newspapers, magazines, books	Not Given	Machine Printed	300 dpi	Not Given	touching characters and heavily printed characters.
Lehal [40]	Multiple sources	31 pages	Not Given	300dpi	Camera	broken characters and heavily printed characters.
Mangla and Kaur [36]	Not Given	120 words	Handwritten	200dpi	Scanner	broken characters and touching characters
Jindal et al. [41]	books and magazines	100 documents	Machine Printed	300 dpi	Scanner	touching characters

segmenting the horizontally overlapping lines and linking the tiny size strips (containing just the lower/upper zone) to their corresponding text lines. One hundred deteriorated printed papers for Gurmukhi text with roughly 6000 touching

characters were collected to test and analyze the middle zone segmentation. All of the zones' categories were evaluated, and segmentation was done based on them. The middle zone was segmented using vertical or horizontal point projection, while the top zone was segmented using structural characteristics. The segmentation and recognition phase focused on the top and middle zones, achieving 76 percent and 77 percent recognition accuracy, respectively.

Figure 6 depicts the performance of existing Gurmukhi recognition systems. Because all suggested methods were evaluated on a small dataset, these findings demonstrate that their performance is more than 90%. Furthermore, these systems handle a variety of degradations such as touched, damaged, or densely printed characters.

The preceding statistics demonstrate that Lehal [40] attained a maximum accuracy of 98 percent. They simply used the camera to grab 31 pages from various sources to test the suggested method. With a short collection of data and in the presence of different degradations, the other systems obtained 96 percent [41], 95 percent [36], and 91.4 percent [37] with a small set of data and in the presence of different degradations, as shown in table .

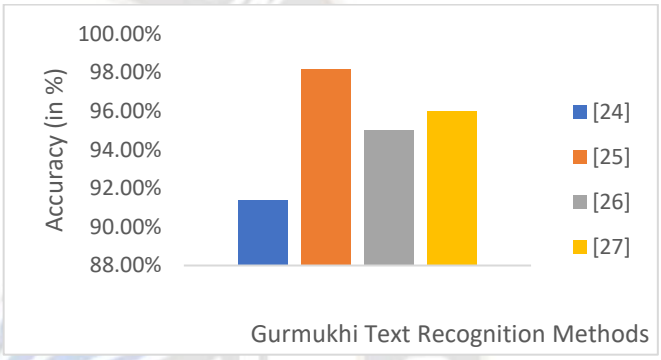


Figure 5. Performance of Gurmukhi Text Recognition Systems.

Table 1 gives details of degradations and presents the source of data, its size, type, resolution, and device used for capturing the data samples. It is clear from the anatomy demonstrated in fig 7 that minimal work is done on the typewritten text compared to Machine Printed and Handwritten Text.

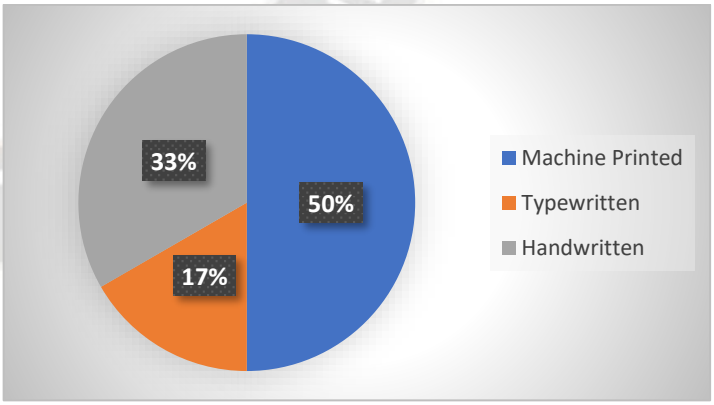


Figure 6. Work was done on Different types of text (Gurmukhi Language)

The effort done on machine-printed text is 34 percent and 66 percent higher than the work done on handwritten and typewritten material, respectively. Even though machine-printed text is quite common nowadays, the ancient literature discovered is typewritten. As a result, typewritten text recognition technologies must be developed.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

2. Deep Learning Based Text Recognition Systems

The specifics of contemporary methods to text recognition systems are discussed in this section. These methods are well-known nowadays and aid in improving the performance of text recognition systems.

Deep Learning is a cutting-edge method that is utilized in a variety of systems. Text Recognition Systems also used deep understanding to recognize text from diverse types of documents. A CNN-based method based on LSTM networks was recently presented [43]. They created a word-level categorization model and used the results to rebuild the word. The suggested method was successfully tested on over 10,000 words from the IAM Handwriting Dataset. For the handwritten dataset, Ingle et al. [44] proposed another LSTM-based method. They presented a line recognition model without recurrent connections in this paper. They tested their model using an extensive dataset and found that this method produced superior outcomes. They classified the handwritten and printed samples and discovered that 98.9% of printed data could be identified, whereas only 59.6% of handwritten data could be recognized. CRConvNet [45] and RCNN [46] were used to recognize characters in multimedia documents, while CRConvNet [45] was utilized to detect printed characters.

Furthermore, CNN-based OCR was created for the Chinese language [47] and showed promising results. Attigeri [48] introduced another recognition method for handwritten data, in which characters were identified using a neural network. They utilized 4400 characters for training their system, and performance was tested using 440 characters. The results represented an accuracy of 85.5% for the recognition of the English characters.

Text detection systems are also common in multimedia data, and they're employed for a variety of things. Number plate extraction, for example, is a high-demand service in today's modern world that may be used in a variety of sectors. Garg et al. [49] developed an ANN-based OCR that identified and retrieved characters from an image sample of printed characters on number plates, taking this into account. This research found that noise significantly influences system performance; thus, data must be cleaned or pre-processed before processing. Furthermore, Mainkar et al. [50] presented feature-based OCR. This approach suggests an Android app that is tested on both handwritten and printed text data. When compared to handwritten text, printed text is more accurate.

The performance of different contemporary approaches using a different types of data has been analyzed based on various parameters and is given in table 2. The outcomes of the suggested deep learning-based techniques are shown in the table above. The following findings show that applying current techniques improves the text recognition system's performance without focusing on a particular sort of data. Handwritten, printed, and a few typewritten samples were included in the aforementioned suggested systems, and their performance was exceptional.

TABLE 2: PERFORMANCE OF CONTEMPORARY APPROACHES

Author	***Type of Dataset***	***Modern Approach***	***Performance***
Naz et al. [25]	Typewritten	MD-LSTM -RNN	Recognition Rate: 96.40%.
Naz et al. [26]	Printed Text	MD-LSTM -RNN	Accuracy: 98%
Naz et al. [27]	Printed Text	MDLSTM	Accuracy: 98.12%
Manchala *et al.* [1]	Handwritten	CNN	Accuracy: Training: 38% Validation: 33%
Ingle et al. [44]	Handwritten and Printed	LSTM	Accuracy: Handwritten: 59.6% Printed: 98.9%
Pradeep [48]	Handwritten	Neural Network	Accuracy: 90.19%
Garg et al. [49]	Printed	ANN	Accuracy: 95%
He [46]	Printed	CRNN	Accuracy: 77%
Mainkar et al. [50]	Handwritten	OCR Engine	Accuracy: 90%
Yadav et al. [45]	Printed	pretrained CRConvNet	F-Measure: 0.85
Puri and Singh [30]	Handwritten	SVM	Accuracy: 98.35%
Thilagavathi et al. [31]	Handwritten	CNN	-
Weng and Xia [47]	Handwritten	CNN	Accuracy 93.3%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

III. PERFORMANCE OF DIFFERENT TEXT RECOGNITION SYSTEMS

The performance of an existing text recognition system in several languages such as English, Spanish, Bangla, and others was assessed based on accuracy and other criteria.

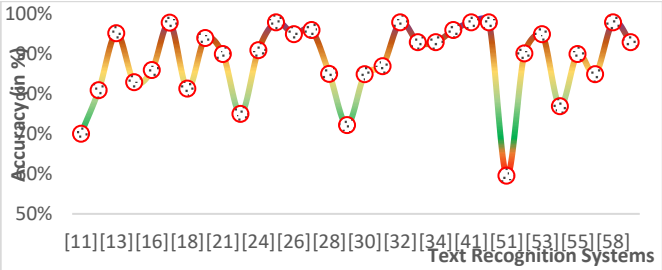


Figure 7. Performance of different Text Recognition systems.

The majority of the systems have greater than 80% accuracy, and all of these systems were created for different languages, as shown in table 4. In this table, the methods used in each phase of recognition are deliberated, along with the details of the dataset and performance. In Figure 7, the performance of various Text Recognition Systems is assessed only on the basis of accuracy. This statistic indicates the accuracy of various text recognition systems, which is greater than 80%. Some of the works even achieve a 90% accuracy rate, sufficient for any text recognition system.

IV. CHALLENGES IN TEXT RECOGNITION SYSTEMS

The current scenario indicates that, even in the face of degradations, the performance of various methods was fairly outstanding for various types of data. The performance specifics for some of the modern techniques are provided in the preceding section. This section discusses the shortcomings of the existing Text Recognition system and provides alternative techniques and their results. The challenges are divided into two categories: (a) Depending on Degradations, and (b) Depending on the availability of the type of data.

(a) Degraded Data Samples

The text recognition systems use different types of sample documents are affected by the specific degradations defined in table 3. These degradations were caused due to the following reasons:

- There are both online and offline versions of handwritten papers. The ink of the pen might also be one of the reasons for such degradations in offline documents that are pen-paper based. A light pen, for example, might result in fractured characters, but a strong ink pen can result in touching or densely written issues. Furthermore, various scanners are utilized to gather these offline samples. The quality of the sample is also affected by the scanning instrument. On the other hand, online handwritten examples are subject to skewness, heavy printing, and other issues dependent on the user.
- Some of these degradations are evident in machine-printed documents due to the printing machinery. Because of the abundant ink, these papers are prone to touching characters and highly printed characters, whereas broken characters are caused by sparse ink.
- Another cause of data deterioration is the equipment that was used to record the data. For example, pictures taken with a camera may be impacted by many types of sounds as a result of internal or external factors.
- The writing style of the users in the case of handwritten samples can also be one of the reasons for degradation.

Typewritten documents have numerous challenges already discussed in section 4.

The table 3 shows the degradations detected in the samples from multiple text recognition systems designed for various languages. As shown in fig 8, the amount of work done on touching character degradation is almost double that of broken, heavy printed characters and four times that of skewness.

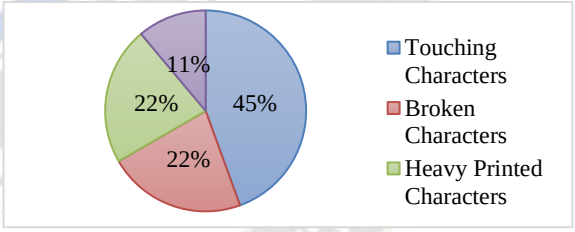


Figure 8. Work was done on Different types of degradation.

(a) Samples Availability

For Text Recognition Systems, there are three sorts of examples available: I Handwritten, (ii) Typewritten, and (iii) Machine Printed. All of these categories have distinct kinds of deterioration, as mentioned in the preceding section. The availability of samples will be examined in this section based on previous research. Figure 9 depicts the work done on various types of data. It has been discovered that handwritten text-based recognition systems are more numerous than machine-printed and typewritten text recognition systems. It is also clear from the interpretation that the machine printed text-based recognition systems are limited.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

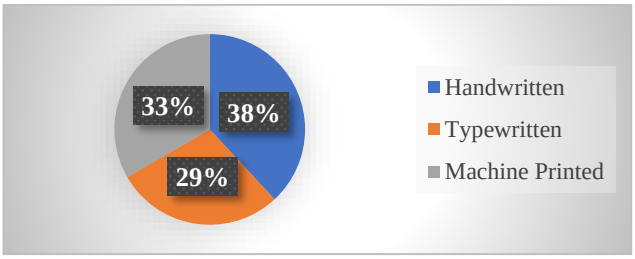


Figure 9. Work done on the different types of Data

Figure 9 depicts the researchers' work on several types of data. Each piece of data pertains to a distinct language, as shown in the accompanying chart. Figure 10 represented the work done in different languages for handwritten, typewritten andmachine-printed data. It is clear from the figure that most of the work is done for the English Language for each type of data as English is the universal language. The work done on the other language is very less Ecompared to English as analyzed in this paper from the works considered in this work of the past few years.

V. CONCLUSION AND FUTURE SCOPE

After This article investigates and analyses text recognition systems, which are widely used NLP systems for digitization. This study also goes through various degradations seen in the dataset used to evaluate and compare various text recognition algorithms. However, the Gurmukhi language received the most attention, and it was discovered that the Gurmukhi Typewritten language had a single recognition mechanism. The article also analysed the accuracy of various text recognition systems. In comparison to other languages, the English Language was given the highest weight in the study, and numerous recognition algorithms were created for the English Language. Even though the typewritten recognition system requires to recover old datasets, it is evident from anatomy that relatively little work is done on this compared to Machine Printed and Handwritten Text Furthermore, the examination of the work done on the various forms of deterioration is examined. It depicts that the work on touching character degradation is nearly double to broken, heavy printed characters and four times greater than skewness.

TABLE 3: DIFFERENT TYPES OF DEGRADATIONS

Author	***Language***	***Degradations***
Jindal et al. [35]	Gurmukhi	touching characters
Jindal et al. [39]	Gurmukhi	broken characters, touching characters, and heavily printed characters.
Jindal et al. [37]	Gurmukhi	touching characters and heavily printed characters.
Lehal [40]	Gurmukhi	broken characters and heavily printed characters.
Mangla and Kaur [36]	Gurmukhi	broken characters and touching characters
Jindal et al. [41]	Gurmukhi	touching characters
Boudraa et al. [52]	English	blur, noise, and back to front interference problems
Ahn et al. [53]	German, English, and Spanish	Touching lines, skewness, and structure noise
Sehad et al. [54]	English	ink degradation, ink bleed-through degradation, and poor contrast
Garain and Chaudhuri [55]	Bangla and Devanagari	touching characters
Ceniza et al. [56]	English and Filipino	Broken Characters, Background Noise, Heavy Printed
Rehman [57]	English	Touching Characters
Bannigidad and Gudada [56]	Kannada	degraded historical

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

TABLE 4: DIFFERENT TEXT RECOGNITION SYSTEMS: METHODS AND PERFORMANCE

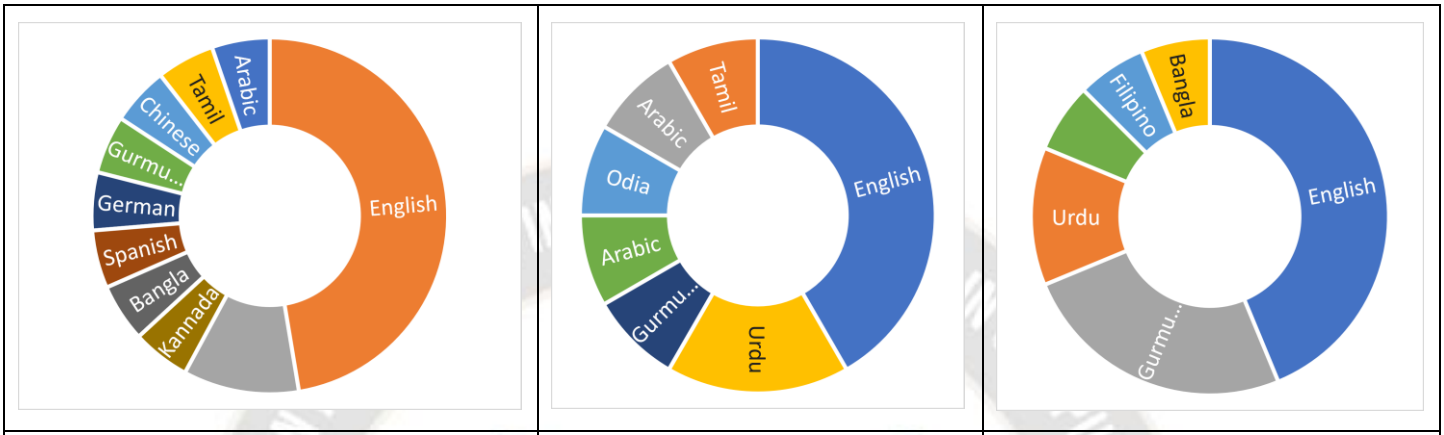
Author	Language	Type of Dataset	Dataset Size	Pre-processing	Segmentation	Feature Extraction	Classification/Recognition	Performance
Pletschacher et al. [20]	English	Typewritten-Historical	643 images	Binarization	Glyph Level Segmentation	geometrical characteristics, density information, and localised features	semi-supervised clustering	F-measure – (ML constraints:0.701, CL constraints:0.59)
Abubacker and Gandhi [21]	Tamil	Typewritten	Not Given	Sigmoid function, horizontal and vertical projection-profiles	Matra / Extension segmentation algorithm	Structural and Statistical Features	DBN Classifier	Accuracy: 81%
Cao et al. [22]	Arabic	Handwritten and Typewritten	8000 images	Gabor Filter	NA	image intensity percentile features, angles and correlation features	HMM	Error Rate: 4.75%
Vučković and Arizanović [23]	English	Machine-typed and Machine-printed	74 images	Skew Correction, Image Filtering	the sliding window-based method with vertical sliding	NA	NA	Min Processing Time-55.53 Max Processing Time-408.02
Kumar and Sharma [24]	English	Handwritten and Typewritten	360 characters	Binarization using Otsu algorithm, median filter.	Connected Component Labelling	NA	Template Matching	Accuracy (Typewritten:83%. Handwritten:60%)
Mahmood and Srivastava [51]	Urdu	Typewritten	115 Images	Not Given	Connected component	NA	NA	Precision: 87.36% Recall: 84.75%
Panda and Tripathy [28]	Odia	Typewritten	52-character images	Binarization Skew Correction	Narrow horizontal band	Not Given	Template Matching	Accuracy: 97.87%
Khorsheed [32]	Arabic	Cursive Typewritten	600 A4-size Sheets	Binarization	No Segmentation	Statistical Features	HTK recognition	Recognition Accuracy: 81.35%
Retsinas et al. [34]	English	Typewritten-Historical	10pages, 15000 characters	ASF Filter Binarization	Connected Components	HoG	GMM	Error Rate: 6%
Naz et al. [25]	Urdu	Typewritten	10,000 text lines	Gray Scale Conversion	NA	Automatic Features by RNN	MD-LSTM -RNN	Recognition Rate: 96.40%.
Naz et al. [26]	Urdu	Printed Text	10,000 images	Normalization	NA	Automatic Features by RNN	MD-LSTM -RNN	Accuracy: 98%
Naz et al. [27]	Urdu	Printed Text	10,000 images	Automatic by CNN	NA	CNN	MDLSTM	Accuracy: 98.12%
Manchala et al. [1]	English	Handwritten	1500 forms	Padding and Rotation	Tesseract model	CNN	VGG-19	Accuracy Training: 38% Validation: 33% Testing: 31%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

Ingle et al. [44]	English	Handwritten and Printed	Historic Images: 1,59,326	Line Extraction	NA	HTR Model	LSTM	Accuracy Handwritten: 59.6% Printed: 98.9%
Pradeep [48]	English	Handwritten	4840 characters	Noise Reduction Binarization Edge Detection Dilation and filling	External and Internal Segmentation	NA	Neural Network	Accuracy: 90.19%
Garg et al. [49]	English	Printed	40 Images	Median Filter	Line, Word, and Character	Connected Pixels	ANN	Accuracy: 95%
He [46]	English	Printed	Image Data	Automatic	Automatic	Automatic	CRNN	Accuracy: 77%
Mainkar et al. [50]	English	Handwritten	221 characters	greyscale conversion, binarization, thinning, skewing and normalization	Line Segmentation	slant angle, height, curves	OCR Engine	Accuracy: 90%
Yadav et al. [45]	English	Printed	360	Automatic using ConvNet	global threshold and watershed segmentation	diagonal base feature learning	pretrained CRConvNet	F-Measure: 0.85
Puri and Singh [30]	Devanagari	Handwritten	60 documents	Outlier removal and Skew Correction	HPP and VPP	geometric based features	SVM	Accuracy: 98.35%
Thilagavathi et al. [31]	Tamil	Handwritten	82934	Automatic	Automatic	Automatic	CNN	-
Weng and Xia [47]	Chinese	Handwritten	50000	Automatic	Automatic	Automatic	CNN	Accuracy 93.3%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8008
***Article Received**: 18 May 2023 **Revised**: 25 July 2023 **Accepted**: 12 August 2023*

Despite the availability of various text recognition systems, there is a requirement for an efficient typewritten text recognition system that considers the most degradations. A special emphasis might be placed on languages that are rarely considered for any material. One of the market demands, for example, is a Typewritten Text Recognition System for Gurmukhi Language. As a result, the primary focus in the future will be on developing an effective text recognition system for Typewritten Gurmukhi Text. Furthermore, different issues discussed in the paper might be explored in the future for improving existing text recognition systems, particularly for typewritten samples.



(a) Handwritten (b) Typewritten (c) Machine Printed

Figure 10. Work done on different Languages

REFERENCES

[1] Y. Manchala, J. Kinthali, K. Kotha, K. S. Kumar, and J. Jayalaxmi, "Handwritten Text Recognition using Deep Learning with TensorFlow," Int. J. Eng. Res., vol. V9, no. 05, pp. 594–600, 2020, doi: 10.17577/ijertv9is050534.

[2] S. IMPEDOVO, L. OTTAVIANO, and S. OCCHINEGRO, "OPTICAL CHARACTER RECOGNITION — A SURVEY," Int. J. Pattern Recognit. Artif. Intell., vol. 05, no. 01n02, pp. 1–24, Jun. 1991, doi: 10.1142/S0218001491000041.

[3] V. . Govindan and A. . Shivaprasad, "Character recognition — A review," Pattern Recognit., vol. 23, no. 7, pp. 671–683, Jan. 1990, doi: 10.1016/0031-3203(90)90091-X.

[4] N. Arica and F. T. Yarman-Vural, "Optical character recognition for cursive handwriting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 6, pp. 801–813, Jun. 2002, doi: 10.1109/TPAMI.2002.1008386.

[5] N. Islam, Z. Islam, and N. Noor, "A Survey on Optical Character Recognition System," J. Inf. Commun. Technol., vol. 10, no. 2, pp. 1–7, Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.05703.

[6] K. Dutta, P. Krishnan, M. Mathew, and C. V. Jawahar, "Improving CNN-RNN Hybrid Networks for Handwriting Recognition," in 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Aug. 2018, pp. 80–85, doi: 10.1109/ICFHR-2018.2018.00023.

[7] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," IEEE Access, vol. 8, pp. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

[8] N. Sankaran and C. V Jawahar, "Recognition of printed Devanagari text using BLSTM Neural Network," in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 1–7.

[9] A. Ray, S. Rajeswar, and S. Chaudhury, "Text recognition using deep BLSTM networks," in 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), Jan. 2015, pp. 1–6, doi: 10.1109/ICAPR.2015.7050699.

[10] M. Z. Alom, P. Sidike, T. M. Taha, and V. K. Asari, "Handwritten Bangla Digit Recognition Using Deep Learning," Comput. Vis. Pattern Recognit., pp. 1–9, May 2017, [Online]. Available: http://arxiv.org/abs/1705.02680.

[11] B. Polaiah, N. S. S. T. Velpuri, G. K. Pandala, S. L. R. S. Polavarapu, and P. R. Kumari, "Handwritten text recognition using machine learning techniques in application of NLP.," Int. J. Innov. Technol. Explor. Enginnering, pp. 1394–1397, 2019.

[12] Zecheng Xie, Zenghui Sun, Lianwen Jin, Ziyong Feng, and Shuye Zhang, "Fully convolutional recurrent network for handwritten Chinese text recognition," in 2016 23rd International Conference on Pattern Recognition (ICPR), Dec. 2016, pp. 4011–4016, doi: 10.1109/ICPR.2016.7900261.

[13] R. Messina and J. Louradour, "Segmentation-free handwritten Chinese text recognition with LSTM-RNN," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Aug. 2015, pp. 171–175, doi: 10.1109/ICDAR.2015.7333746.

[14] S. Bhowmik, R. Sarkar, B. Das, and D. Doermann, "GiB : A ${G}$ ame Theory ${I}$ nspired ${B}$ inarization Technique for Degraded Document Images," IEEE Trans. Image Process., vol. 28, no. 3, pp. 1443–1455, Mar. 2019, doi:

10.1109/TIP.2018.2878959.

[15] A. Rehman and T. Saba, "Neural networks for document image preprocessing: state of the art," Artif. Intell. Rev., vol. 42, no. 2, pp. 253–273, Aug. 2014, doi: 10.1007/s10462-012-9337-z.

[16] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," IEEE Trans. Syst. Man Cybern. Part C (Applications Rev., vol. 31, no. 2, pp. 216–233, May 2001, doi: 10.1109/5326.941845.

[17] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 7, pp. 690–706, Jul. 1996, doi: 10.1109/34.506792.

[18] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Mar. 2016, pp. 2264–2268, doi: 10.1109/WiSPNET.2016.7566545.

[19] S. Ding, H. Zhu, W. Jia, and C. Su, "A survey on feature extraction for pattern recognition," Artif. Intell. Rev., vol. 37, no. 3, pp. 169–180, Mar. 2012, doi: 10.1007/s10462-011-9225-y.

[20] S. Pletschacher, J. Hu, and A. Antonacopoulos, "A New Framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering," in 2009 10th International Conference on Document Analysis and Recognition, 2009, pp. 506–510, doi: 10.1109/ICDAR.2009.267.

[21] N. F. Abubacker and R. I. Gandhi, "An extended method for recognition of broken typewritten characters special reference to tamil script," in 2011 IEEE Conference on Open Systems, Sep. 2011, pp. 214–219, doi: 10.1109/ICOS.2011.6079265.

[22] H. Cao, R. Prasad, and P. Natarajan, "Handwritten and Typewritten Text Identification and Recognition Using Hidden Markov Models," in 2011 International Conference on Document Analysis and Recognition, Sep. 2011, pp. 744–748, doi: 10.1109/ICDAR.2011.155.

[23] V. Vučković and B. Arizanović, "Efficient character segmentation approach for machine-typed documents," Expert Syst. Appl., vol. 80, pp. 210–231, Sep. 2017, doi: 10.1016/j.eswa.2017.03.027.

[24] S. Kumar and P. Sharma, "Offline Handwritten & Typewritten Character Recognition using Template Matching," Int. J. Comput. Sci. Eng. Technol., vol. 4, no. 06, pp. 818–825, 2013, [Online]. Available: http://www.ijcset.com/docs/IJCSET13-04-06-085.pdf.

[25] S. Naz et al., "Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks," Neurocomputing, vol. 177, pp. 228–241, Feb. 2016, doi: 10.1016/j.neucom.2015.11.030.

[26] S. Naz, A. I. Umar, R. Ahmed, M. I. Razzak, S. F. Rashid, and F. Shafait, "Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks," Springerplus, vol. 5, no. 1, p. 2010, Dec. 2016, doi: 10.1186/s40064-016-3442-4.

[27] S. Naz et al., "Urdu Nastaliq recognition using convolutional–recursive deep learning," Neurocomputing, vol. 243, pp. 80–87, Jun. 2017, doi: 10.1016/j.neucom.2017.02.081.

[28] S. R. Panda and J. Tripathy, "Odia Offline Typewritten Character Recognition using Template Matching with Unicode Mapping," in 2015 International Symposium on Advanced Computing and Communication (ISACC), Sep. 2015, pp. 109–115, doi: 10.1109/ISACC.2015.7377325.

[29] M. Sonkusare, R. Gupta, and A. Moghe, "A Review on Handwritten Devanagari Character Recognition," EasyChair, pp. 1–6, 2019.

[30] S. Puri and S. P. Singh, "An efficient Devanagari character classification in printed and handwritten documents using SVM," Procedia Comput. Sci., vol. 152, pp. 111–121, 2019, doi: 10.1016/j.procs.2019.05.033.

[31] G. Thilagavathi, G. Lavanya, and N. K. Karthikeyan, "Tamil handwritten character recognition using artificial neural network," Int. J. Sci. Technol. Res., vol. 8, no. 12, pp. 1611–1616, 2019.

[32] M. S. Khorsheed, "Recognizing Cursive Typewritten Text Using Segmentation-Free System," Sci. World J., vol. 2015, pp. 1–7, 2015, doi: 10.1155/2015/818432.

[33] M. S. Khorsheed, "Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)," Pattern Recognit. Lett., vol. 28, no. 12, pp. 1563–1571, Sep. 2007, doi: 10.1016/j.patrec.2007.03.014.

[34] G. Retsinas, B. Gatos, A. Antonacopoulos, G. Louloudis, and N. Stamatopoulos, "Historical Typewritten Document Recognition Using Minimal User Interaction," in Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, Aug. 2015, pp. 31–38, doi: 10.1145/2809544.2809559.

[35] M. K. Jindal, G. S. Lehal, and R. K. Sharma, "A Study of Touching Characters in Degraded Gurmukhi Text," Eng. Technol., vol. 4, no. February, pp. 121–124, 2005.

[36] P. Mangla and H. Kaur, "An end detection algorithm for segmentation of broken and touching characters in handwritten Gurumukhi word," in Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, Oct. 2014, pp. 1–4, doi: 10.1109/ICRITO.2014.7014740.

[37] M. K. Jindal, R. K. Sharma, and G. S. Lehal, "Structural Features for Recognizing Degraded Printed Gurmukhi Script," in Fifth International Conference on Information Technology: New Generations (itng 2008), Apr. 2008, pp. 668–673, doi: 10.1109/ITNG.2008.223.

[38] J. S. Madan, R. Sidhu, and D. V. Sharma, "Development of a generic structural feature extraction method for printed Gurumukhi and similar scripts," in 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1–7.

[39] M. K. Jindal, R. K. Sharma, and G. S. Lehal, "A Study of Different Kinds of Degradation in Printed Gurmukhi Script," in 2007 International Conference on Computing: Theory and Applications (ICCTA'07), Mar. 2007, pp. 538–544, doi: 10.1109/ICCTA.2007.19.

[40] G. S. Lehal, "Optical character recognition of Gurmukhi script using multiple classifiers," in Proceedings of the International Workshop on Multilingual OCR - MOCR '09,

2009, pp. 1–7, doi: 10.1145/1577802.1577810.

[41] M. K. Jindal, G. S. Lehal, and R. K. Sharma, "Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script," Int. J. Signal Process., vol. 2, no. 4, pp. 258–267, 2008.

[42] R. J. Shah and T. V. Ratanpara, "Challenges of broken characters in character segmentation method for Gujarati printed documents," in 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2015, pp. 1–5, doi: 10.1109/ICIIECS.2015.7193263.

[43] A. Singh, Y. Ghasemi, H. Jeong, M. Kim, and A. Johnson, "A comparative evaluation of the wearable augmented reality-based data presentation interface and traditional methods for data entry tasks," Int. J. Ind. Ergon., vol. 86, p. 103190, Nov. 2021, doi: 10.1016/j.ergon.2021.103190.

[44] R. R. Ingle, Y. Fujii, T. Deselaers, J. Baccash, and A. C. Popat, "A Scalable Handwritten Text Recognition System," in 2019 International Conference on Document Analysis and Recognition (ICDAR), Sep. 2019, pp. 17–24, doi: 10.1109/ICDAR.2019.00013.

[45] U. Yadav, S. Verma, D. K. Xaxa, and C. Mahobiya, "A deep learning based character recognition system from multimedia document," in 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Apr. 2017, pp. 1–7, doi: 10.1109/IPACT.2017.8245200.

[46] Y. He, "Research on Text Detection and Recognition Based on OCR Recognition Technology," in 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), Sep. 2020, pp. 132–140, doi: 10.1109/ICISCAE51034.2020.9236870.

[47] Y. Weng and C. Xia, "A New Deep Learning-Based Handwritten Character Recognition System on Mobile Computing Devices," Mob. Networks Appl., vol. 25, no. 2, pp. 402–411, Apr. 2020, doi: 10.1007/s11036-019-01243-5.

[48] J. Pradeep, E. Srinivasan, and S. Himavathi, "Neural network based handwritten character recognition system without feature extraction," in 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), Mar. 2011, pp. 40–44, doi: 10.1109/ICCCET.2011.5762513.

[49] S. Garg, K. Kumar, N. Prabhakar, A. Ratan, and A. Trivedi, "Optical Character Recognition using Artificial Intelligence," Int. J. Comput. Appl., vol. 179, no. 31, pp. 14–20, Apr. 2018, doi: 10.5120/ijca2018916390.

[50] V. V. Mainkar, J. A. Katkar, A. B. Upade, and P. R. Pednekar, "Handwritten Character Recognition to Obtain Editable Text," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Jul. 2020, pp. 599–602, doi: 10.1109/ICESC48915.2020.9155786.

[51] A. Mahmood and A. Srivastava, "A Novel Segmentation Technique for Urdu Type-Written Text," in 2018 Recent Advances on Engineering, Technology and Computational Sciences (RAETCS), Feb. 2018, pp. 1–5, doi: 10.1109/RAETCS.2018.8443958.

[52] O. Boudraa, W. K. Hidouci, and D. Michelucci, "A robust multi stage technique for image binarization of degraded historical documents," in 2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B), Oct. 2017, pp. 1–6, doi: 10.1109/ICEE-B.2017.8192044.

[53] B. Ahn, J. Ryu, H. Il Koo, and N. I. Cho, "Textline detection in degraded historical document images," EURASIP J. Image Video Process., vol. 2017, no. 1, pp. 82–90, Dec. 2017, doi: 10.1186/s13640-017-0229-7.

[54] A. Sehad, Y. Chibani, R. Hedjam, and M. Cheriet, "Gabor filter-based texture for ancient degraded document image binarization," Pattern Anal. Appl., vol. 22, no. 1, pp. 1–22, Feb. 2019, doi: 10.1007/s10044-018-0747-7.

[55] U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed devnagari and bangla scripts using fuzzy multifactorial analysis," IEEE Trans. Syst. Man Cybern. Part C (Applications Rev., vol. 32, no. 4, pp. 449–459, Nov. 2002, doi: 10.1109/TSMCC.2002.807272.

[56] A. M. Ceniza, T. K. B. Archival, and K. V. Bongo, "Mobile Application for Recognizing Text in Degraded Document Images Using Optical Character Recognition with Adaptive Document Image Binarization," J. Image Graph., vol. 6, no. 1, pp. 44–47, 2018, doi: 10.18178/joig.6.1.44-47.

[57] A. Rehman, "Offline touched cursive script segmentation based on pixel intensity analysis: Character segmentation based on pixel intensity analysis," in 2017 Twelfth International Conference on Digital Information Management (ICDIM), Sep. 2017, pp. 324–327, doi: 10.1109/ICDIM.2017.8244641.