**TITLE OF PROJECT**

Optical Character Recognition Software(OCR)

**Project Report**

*Submitted in the partial fulfillment for the award of the degree of*

# BACHELOR OF ENGINEERING

**IN**

**BIG DATA ANALYTICS**

**Submitted by:** PRAJWAL UNIYAL - 19BCS3875

AMRIT SINGH - 20BCS8501

Alok – 19BCS3900

GAVIN – 20BCS8502

# IN THE SUPERVISION OF:

# Mr. SUHAIL JAVED QURAISH



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING APEX INSTITUE OF TECHNOLOGY**

**CHANDIGARH UNIVERSITY, GHARUAN, MOHALI 140413, PUNJAB**

# DECLARATION

I, Amit Singh , Prajwal Uniyal , Alok , Gavin student of Big data analytics, Department of Computer Science and Engineering, Apex Institute of Technology, Chandigarh University, Punjab, hereby declare that the work

presented in this Project Work entitled ‘**OCR’** is the outcome of our own bona fide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. It contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

**(CandidateName):**

**PRAJWAL UNIYAL**

**AMRITSINGH**

**ALOK**

**GAVIN**

**CandidateUID:**

**19BCS3875**

**20BCS8501**

**19BCS3900**

**20BCS8502**

**Date-25-April-2022**

**Place-Chandigarh**

**Content:**

|  |  |
| --- | --- |
| Title Page | i |
| Declaration of the Student | ii |
| Abstract | iii |
| Acknowledgement | iv |
| List of Figures | v |
| List of Tables (optional) | vi |
| Timeline / Gantt Chart | vii |
| **INTRODUCTION** | **1** |
| 1.1 Problem Definition | 1 |
| 1.2 Project Overview/Specifications\* (page-1 and 3) | 2 |
| 1.3 Hardware Specification | 3 |
| 1.4 Software Specification | 4 |
|  | 4 |
| … |  |
| **LITERATURE SURVEY** | **5** |
| 2.1 Existing System | 5 |
| 2.2 Proposed System | 6 |
| 2.3 Feasibility Study\* (page-4) | 7 |
| **PROBLEM FORMULATION** | 9 |
| **OBJECTIVES** | 16 |
| **METHODOLOGY** | 18 |
| **CONCLUSIONS AND DISCUSSION** | 19 |
| **REFERENCES** | 22 |

**OCR**

**SYSTEM**

# INTRODUCTION

In the running world, there's growing demand for the software systems to acknowledge characters in ADP system when information is scanned throughpaper

documents as we know that we've got number of newspapers and books which are in printed format associated with different subjects. nowadays there's an enormous demand in “storing the knowledge available in these paper documents in to a computer memory disk and so later reusing this information by searching process”. One simple thanks to store information in these paper documents in to ADP system is to first scan the documents so store them as images. But to reuse this information it's very difficult to read the individual contents and searching the contents form these documents line-by-line and word- by-word. The reason for this difficulty is that the font characteristics of the characters in paper documents are different to font of the characters in system. As a result, computer is unable to recognize the characters while reading them. this idea of storing the contents of paper documents in hardware place so reading and searching the content is termed Document processing. Sometimes during this document processing we want to process the information that's associated with languages aside from country within the world. For this document processing we'd like a code called Character recognization system. This process is additionally called Document image analysis(DIA).

The conversion of paper documents in to electronic format is an on- going task in many of the organizations particularly in Research and Development (R&D) area, in large business enterprises, in government institutions, so on. From our problem statement we can introduce the necessity of Optical Character Recognition in mobile electronic devices such as cell phones, digital cameras to acquire images and recognize them as a part of face recognition andvalidation.

Optical character recognition (OCR) is a system that converts input text into machine-encoded format . Today, OCR is helping not only in digitizing the handwritten medieval manuscripts , but also helps in converting the typewritten documents into digital form . This has made the retrieval of the required information easier as one doesn’t have to go through the piles of documents and files to search the required information. Organizations are satisfying the needs of digital preservation of historic data , law documents , educational persistence etc.

An OCR system depends mainly, on the extraction of features and discrimination/classification of these features (based on patterns). Handwritten OCR have received increasing attention as a subfield of OCR. It is further categorized into offline system ,  and online system  based on input data. The offline system is a static system in which input data is in the form of scanned images while in online systems nature of input is more dynamic and is based on the movement of pen tip having certain velocity, projection angle, position and locus point. Therefore, an online system is considered more complex and advance, as it resolves the overlapping problem of input data that is present in the offline system.

One of the earliest OCR systems was developed in the 1940s, with the advancement in the technology over the time, the system became more robust to deal with both printed, and handwritten characters and this led to the commercial availability of the OCR machines. In 1965, advance reading machine “IBM 1287” was introduced at the “world fair” in New York . This was the first-ever optical reader, which was capable of reading handwritten numbers. During the 1970s, researchers focused on the improvement of response time and performance of the OCR system.

The next two decades from 1980 till 2000, the software system of OCR was developed and deployed in educational institutes, census OCR  and for recognition of stamped characters on metallic bar . In the early 2000s, binarization techniques were introduced to preserve historical documents in digital form and provide researchers with access to these documents . Some of the challenges of binarization of historical documents were the use of nonstandard fonts, printing noise and spacing. In mid of 2000, multiple applications were introduced that were helpful for differently-abled people. These applications helped these people in developing reading and writing skills.

In the current decade, researchers have worked on different machine learning approaches which include Support Vector Machine (SVM), Random Forests (RF), k Nearest Neighbor (k NN), Decision Tree (DT) , Neural Networks etc. Researchers combined these machine learning techniques with image processing techniques to increase the accuracy of the optical character recognition system. Recently researchers have focused on developing techniques for the digitization of handwritten documents, primarily based on deep learning  approach. This paradigm shift has been sparked due to adaption of cluster computing and GPUs and better performance by deep learning architectures , which includes Recurrent Neural Networks (RNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) networks etc.

This Systematic Literature Review (SLR) serves not only the purpose of presenting literature in the domain of OCR for different languages but also highlight research directions for a new researcher by highlighting weak areas of current OCR systems that need further investigation.

This article is organized as follows. Section II discusses review methodology employed in this article. Review methodology section includes review protocol, inclusion and exclusion criteria, search strategy, selection process, quality assessment criteria and metadata synthesis of selected studies. Statistical data from selected studies are presented in Section III. Section IV presents research question and their motivation. Section V discusses different classifications methods which are used for handwritten OCR. This section will also elaborate on structural and statistical models for optical character recognition. Section VI presents different databases (for specific language) which are available for research purpose. Section VII presents research overview of language-specific research in OCR, while Section VIII highlights research trends. Section IX summarizes findings and also highlights gaps in research that need the attention of the research community.

# LITERATURE REVIEW

The scope of our product is Human Character Recognition on grid infrastructure to produce an efficient and advanced software tool for users to form Document Image Analyzing, analyzing texts by reading and identifying characters in research, academic, government and business organizations with an outsized dam written, scanned images. It doesn't matter the dimensions of the documents or the kind of characters within the text, the merchandise recognizes, searches and processes very quickly consistent with the requirements ofnature.

The good thing about proposed system that overcomes the downside of the present system is that it supports multiple functionalities like editing and searching. It also adds benefit by providing heterogeneous characters recognition. Our proposed system is OCR on a grid infrastructure which could be a character recognition system.The multiple functionalities include editing and searching too where because the existing system supports only editing of the document during this context In the running world there's a growing demand for the users to convert the printed documents in to electronic documents for maintaining the safety of their data. Hence the basic OCR system was invented to convert the information available on papers in to computer process able documents, in order that the documents may be editable and reusable. the present system/the previous system of OCR on a grid infrastructure is simply OCR without grid functionality. that's the prevailing system deals with the homogeneous character recognition or character recognition of singlelanguages.

# Literature Review Summary

The basic OCR system was invented to convert the information available on papers in to computer process able documents, in order that the documents may be editable and reusable. the present system/the previous system of OCR on a grid infrastructure is simply OCR without grid functionality. that's the prevailing system deals with the homogeneous character recognition or character recognition of single languages.

# PROBLEM FORMULATION

Developers tend to clone fragments of software during development to save efforts and expedite the event process From the literature review,it's

observed that studies highlightthe requirement of efficient and scalable approach for detecting code clones having software vulnerability. the present techniques don't seem to be able to detect every kind of vulnerable code clones. Different approaches suffer from high false negative rate and not scalable to large software systems thanks to time complexity. So firstly, there's a necessity........................................ Second same subject systems should be accustomed compare the approaches whichdetect

# RESEARCH OBJECTIVES

Our proposed system is OCR on a grid infrastructure which may be a character recognition system that supports recognition of the characters of multiple languages.

This feature is what we call grid infrastructure which eliminates the matter of heterogeneous character recognition

The multiple functionalities include editing and searching too where because the existing system supports only editing of the document during this context

Thus OCR on a grid infrastructure is multi-lingual.

we identify the audience who have an interest with the merchandise and are involved within the implementation of the merchandise either directly or indirectly. As from our research, the OCR system is principally useful in R&D at various scientific organizations, in governmental institutes and in large business organizations, we identify the subsequent as various interested audience in implementing OCR system:- The scientists, the research scholars and also the research fellows in telecommunication institutions have an interest in using OCR system for processing the word document that contains base paper for his or her research. The Librarian to manage the data contents of the older books in building virtual digital library requires use of OCR system. Various sites that vendor e-books have an enormous requirement of this OCRsystem

These suggestions could also be effective and useful for the beginners of the merchandise instead of the regular users like research scholars, librarians and administrators of assorted web-sites. With these suggestions, the user need not waste his time in scrolling the documents up and down, browsing through the online, visiting libraries in search of variousbooks

it might facilitate your if you begin with Wikipedia.com. It helps you to know the essential concept of each keyword you need. First learn from it what's OCR? and the way does it work supported a Grid infrastructure? Now you'll proceed your further reading with the introduction of our product we provided in our documentation.

## Review Methods

As mentioned above, this Systematic Literature Review (SLR) aims to identify and present literature on OCR by formulating research questions and selecting relevant research studies. Thus, in summary, this review aims:

1. To summarize existing research work (machine learning techniques and databases) on different languages of handwritten character recognition systems.
2. To highlight research weakness in order to eliminate them through additional research.
3. To identify new research areas within the domain of OCR.

We will follow the strategies proposed by Kitchenham *et al.* [22]. Following the proposed strategy, in subsequent sub-sections review protocol, inclusion and exclusion criteria, search strategy process, selection process and data extraction and synthesis processes are discussed.

### A. Review Protocol

Following the philosophy, principles and measures of the Systematic Literature Review (SLR) [22], this systematic study was initialized with the development of comprehensive review protocol. This protocol identifies review background, search strategy, data extraction, research questions and quality assessment criteria for the selection of study and data analysis.

The review protocol is what that creates a distinction between an SLR and traditional literature review or narrative review [22]. It also enhances the consistency of the review and reduces the researchers’ biases. This is due to the fact that researchers have to present a search strategy and the criteria for the inclusion of exclusion of any study in the review.

### B. Inclusion and Exclusion Criteria

Setting up an inclusion and exclusion criteria makes sure that only articles that are relevant to study are included. Our criteria include research studies from journals, conferences, symposiums and workshops on the optical character recognition of English, Urdu, Arabic, Persian, Indian and Chinese languages. In this SLR, we considered studies that were published from January 2000 to December 2019.

Our initial search based on the keywords only resulted in 1150 research articles related to handwritten OCRs of different languages (refer Figure 1 for complete overview of the selection process). After a thorough review of the articles, we excluded articles that were not clearly related to a handwritten OCR, but appeared in the search, because of keyword match. Additionally, articles were also excluded based on duplicity, non-availability of full text and if the studies were not related to any of our research questions.

**FIGURE 1.**

### 

### C. Search Strategy

Search strategy comprises of automatic and manual search, as shown in Figure 1. An automatic search helped in identifying primary studies and to achieve a broader perspective. Therefore, we extended the review by the inclusion of additional studies. As recommended by Kitchenham *et al.* [22], the manual search strategy was applied to the references of the studies that are identified after the application of automatic search.

For automatic search, we used standard databases which contain the most relevant research articles. These databases include IEEE Explore, ISI Web of Knowledge, Scopus—Elsevier and Springer. While there is plenty of literature available in the magazine, working papers, newspapers, books and blogs, we did not choose them for this review article as concepts discussed in these sources are not subjected to review process; thus their quality cannot be reliably verified.

General keywords derived from our research questions and the title of the study were used to search for research articles. Our aim was to identify as many relevant articles as possible from the main set of keywords. All possible permutations of Optical character recognition concepts were tried in the search, such as “optical character recognition”, “pattern recognition and OCR”, “pattern matching and OCR” etc.

Once the primary data were obtained by using search strings, the data analysis phase of the obtained research papers began with the intention of considering their relevance to research questions and inclusion and exclusion criteria of the study. After that, a bibliography management tool, i.e. Mendeley, was used for storing all related research articles to be used for referencing purpose. Mendeley also helped in identifying duplicate studies.

A manual search was performed with an automatic search to make sure that we had not missed anything. This was achieved through forward and backwards referencing. Furthermore, for data extraction, all the results were imported into a spreadsheet. Snowballing, which is an iterative process in which references of references are verified to identify more relevant literature, was applied to primary studies in order to extract more relevant primary studies. Set of primary studies post snowball process was then added to Mendeley.

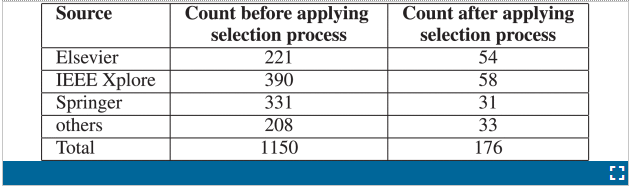
### D. Study Selection Process

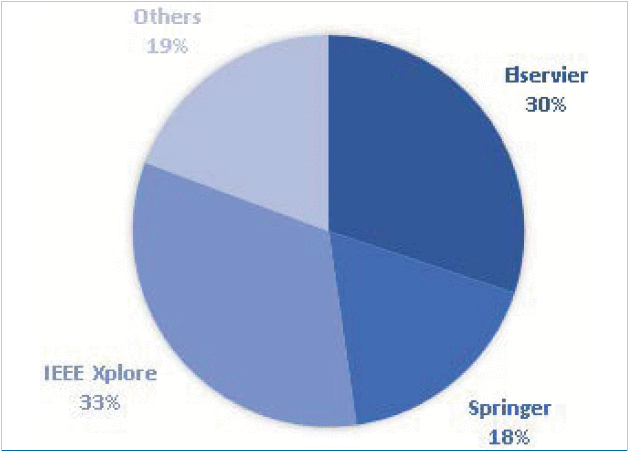
A tollgate approach was adopted for the selection of study [23]. Therefore, after searching keywords in all relevant databases, we extracted 1150 research studies through automatic search. Majority of these 1150 studies, 625 were duplicate studies and were eliminated. Inclusion and exclusion criteria based upon title, abstracts, keywords and the type of publication was applied to the remaining 525 studies. This resulted in the exclusion of 268 studies and leaving 257 studies. In the next stage, the selection criteria were applied, thus further 102 studies were excluded, and we were left with 155 studies.

Once we finished the automatic search stage, we started a manual search procedure to guarantee the exhaustiveness of the search results. We performed screening of the remaining 155 studies and went through the references to check relevant research articles that could have been left search during the automatic search. Manual search added 46 further studies. After adding these studies, pre-final list of 201 primary studies was obtained.

Next and final stage was to apply the quality assessment criteria (QAC) on the pre-final list of 201 studies. Quality assessment criteria were applied at the end as this is the final step through which a final list of studies for SLR was deduced. QAC usually identifies studies whose quality is not helpful in answering the research question. After applying QAC, 25 studies were excluded, and we were left with 176 primary studies. Refer Figure 1 for complete step-by-step overview of selection process.

Table 1 shows the distribution of the primary/selected studies among various publication sources, before and after applying above mentioned selection process. The same is also shown in Figure 2.





.

[Show All](https://ieeexplore.ieee.org/document/9151144/all-figures)

### E. Quality Assessment Criteria

Quality Assessment Criteria (QAC) is based on the principle to make a decision related to the overall quality of the selected set of studies [22]. Following criteria were used to assess the quality of selected studies. This criterion helped us to identify the strength of inferences and helped us in selecting the most relevant research studies for our research.

Quality Assessment criteria questions:

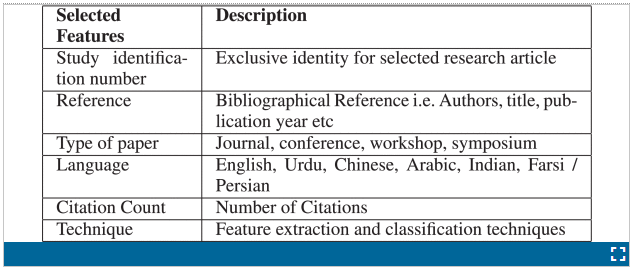
1. Are topics presented in a research paper relevant to the objectives of this review article?
2. Does research study describes the context of the research?
3. Does research article explains the approach and methodology of research with clarity?
4. Is data collection procedure explained If data collection is done in the study?
5. Is the process of data analysis explained with proper examples?

We evaluated 201 selected studies by using the abovementioned quality assessment questions in order to determine the credibility of a particular acknowledged study. These five QA schema is inspired by [23]. The quality of the study was measured depending upon the score of each QA question. Each question was assigned 2 marks, and the study’s quality was considered to be selected if it scored greater than or equal to 5 at the scale of 10. Thus, studies below the score of 5 were not included in the research. Following this criterion, 176 studies were finally selected for this review article (refer to Figure 1 for complete overview of the selection process).

### F. Data Extraction and Synthesis

During this phase, metadata of selected studies (176) was extracted. As stated earlier, we used Mendeley and MS Excel to manage the metadata of these studies. The main objective of this phase was to record the information that was obtained from the initial studies [22]. The data containing study ID (to identify each study), study title, authors, publication year, publishing platform (conference proceedings, journals, etc.), citation count, and the study context (techniques used in the study) were extracted and recorded in an excel sheet. This data was extracted after a thorough analysis of each study to identify the algorithms and techniques proposed by the researchers. This also helped us to classify the studies according to the languages on which the techniques were applied. Table 2 shows the fields of the data extracted from research studies.

**TABLE 2**Extracted Meta-Data Fields of Selected Studies

[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8948470/9151144/sami.t2-3012542-large.gif)

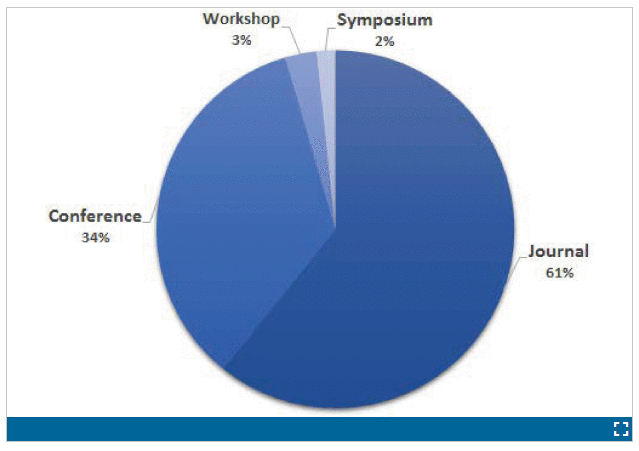
**SECTION III.**

## Statistical Results from Selected Studies

In this section, statistical results of the selected studies are presented with respect to their publication sources, citation count status, temporal view, type of languages and type of research methodologies.

### A. Publication Sources Overview

In this review, most of the included studies are published in reputed journals and leading conferences. Therefore, considering the quality of research studies, we believe that this systematic review can be used as a reference to find latest trends and to highlight research directions for further studies in the domain of handwritten OCR. Figure 3 shows the distribution of studies derived from different publication sources. Majority of included studies (107) were published in research journals (61%), followed by (61) publications in conference articles (34%). Whereas, few (5) articles were published in workshop proceedings and only [(3)](https://ieeexplore.ieee.org/document/" \l "deqn2-deqn3) relevant articles were found to be presented in symposiums.



### B. Research Citations

Citation count was obtained from Google Scholar. Overall, selected studies have good citation count, which shows that the quality of selected studies is worthy of being added in the review and also implies that researchers are actively working in this area of research. As presented in Figure 4, approximately 95% of the selected studies have at least one citation, except the research articles, which are published recently in 2019. Among selected studies, 33 studies have more than 100 citations, 15 studies have been cited between 61–100 times, 25 studies were cited between 33–60 times, 16 studies were cited between 16–30 times and 68 studies were cited between 1 and 15 times. Overall, we predict that selected studies citations will increase further because research articles are constantly being published in this domain.

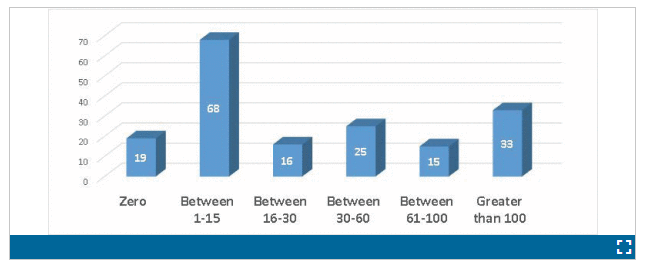
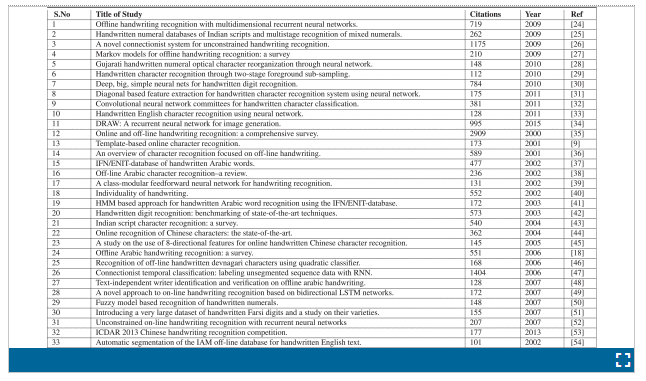


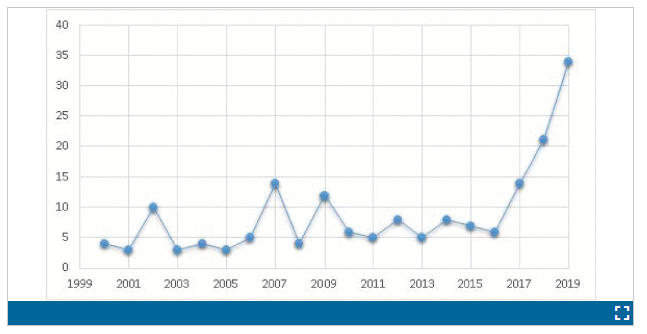
Table 3 provides details of research publications with more than 100 citations each. These articles can be considered to have a strong impact on the researchers working to build robust OCR system.

**TABLE 3**Research Publications With More Than 100 Citations



### C. Temporal View

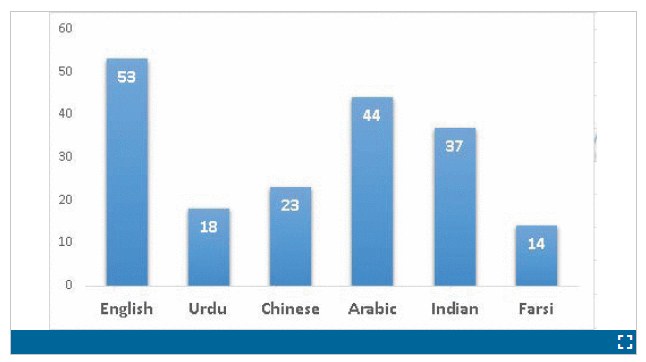
The distribution of number of studies over the period under study (2000 – 2019) can be seen in Figure 5. According to the reference figure, it can be noticed that there is a variation in the publication count through these years. Statistics show sudden increase in a number of publications in the domain of handwritten character recognition in the years 2002, 2007 and 2009. The number of publications remained steady in the remaining years of the 2000s. After 2010 there is again a steady increase in the number publications, i.e. 59 publications in 8 years from 2010-17. During the last two years, we have seen a steep rise in the number of the publication. We found 55 new studies in the last two years as compared 59 studies in the previous 8 years. This is conceivably not surprising since the concept of handwritten character recognition is catching the interest of more researcher because of the advancement of the research work in the fields of deep learning and computer vision. We believe that application areas of handwritten OCRs will further increase in the coming years. This is to be noted that these number of studies only include research articles which are related to our research questions.



Publications over the years. On the y -axis is the number of publications.

### D. Language Specific Research

The distributions/number of selected studies with respect to investigated scripting languages are shown in Figure 6. A total number of selected studies are 176, and out of these 172 studies, the English language has the highest contribution of 53 studies in the domain of handwritten character recognition, 44 studies related to the Arabic language, 37 studies are on the Indian scripts, 23 on the Chinese language, 118 on the Urdu language, while 14 studies were conducted on the Persian language. Some of the selected articles discussed multiple languages.

Number of selected studies with respect to investigated language. Numeric value within bar shows number of selected studies for the given language.

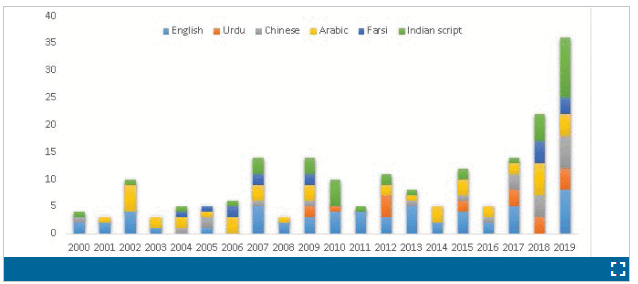


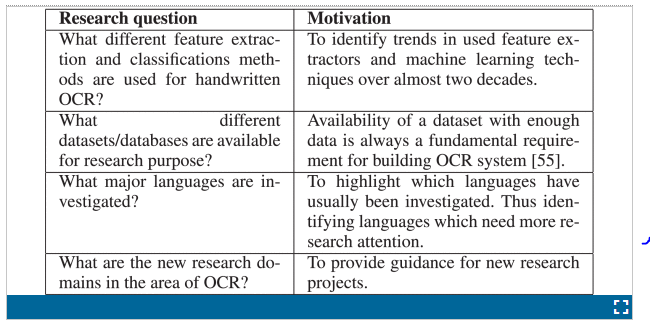
Figure 7 represents publications count each year with respect to language. Reference figure shows compiled temporal view of handwritten OCR researches done in different languages throughout the mentioned era of 2000-2019, in this time period there are certain research articles that cover more than one language of handwritten OCR.

**SECTION IV.**

## Research Questions

Research questions play an important role in a systematic literature review because these questions determine the search queries and keywords that will be used to explore research publications. As discussed above, we chose research questions which not only help seasoned researchers but also to researchers entering in the domain of optical character recognition to understand where the research in this field stands as of today. This review article answers research questions presented in Table 4. Reference table also presents the motivation for each research question.

**TABLE 4**



**SECTION V.**

## Classification Methods of Handwritten OCR

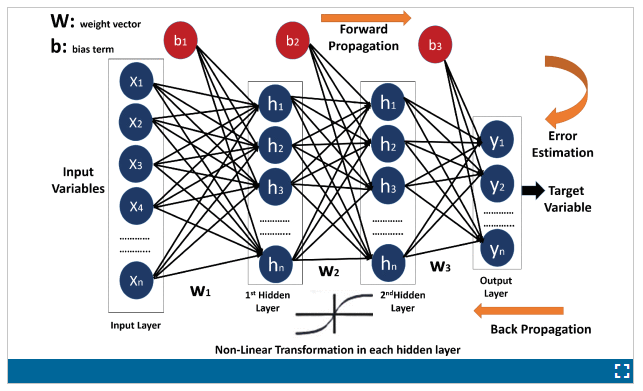
In handwritten OCR an algorithm is trained on a known dataset, and it discovers how to accurately categorize/classify the alphabets and digits. Classification is a process to learn a model on a given input data and map or label it to predefined category or classes [17]. In this section, we have discussed the most prevalent classification techniques in OCR research studies beginning from 2000 till 2019.

### A. Artificial Neural Networks (ANN)

Biological neuron inspired architecture, Artificial Neural Networks (ANN) consists of numerous processing units called neurons . These processing elements (neurons) work together to model given input data and map it to predefined class or label . The main unit in neural networks is nodes (neuron). Weights associated with each node are adjusted to reduce the squared error on training samples in a supervised learning environment (training on labelled samples/data). Figure 8 presents a pictorial representation of Multi-Layer Perceptron (MLP) that consists of three layers, i.e. (input, hidden and output).

**FIGURE 8.**

An architecture of Multilayer Perceptron (MLP) [58].



Feedforward networks/Multi-Layer Perceptron (MLP) achieved renewed interest of research community in the mid 1980s as by that time “Hopfield network” provided the way to understand human memory and calculate the state of a neuron [59]. Initially, the computational complexity of finding weights associated with neurons hindered the application of neural networks. With the advent of deep (many layers) neural architectures, i.e. Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN), neural networks have established itself as one of the best classification technique for recognition tasks including OCR . Refer Sections VIII and IX-B for current and future research trends.

The early implementation of MLP for handwritten OCR was done by Shamsher *et al.* on the .Urdu language. The researchers proposed feed-forward neural network algorithm of MLP (Multi-Layer Perceptrons) . Liu and Suen  used MLP on Farsi and Bangla numerals. One hidden layer was used with the connecting weights estimated by the error backpropagation (BP) algorithm that minimized the squared error criterion. On the other hand, Cirecsan *et al.*  trained five MLPs with two to nine hidden layers and varying numbers of hidden units for the recognition of English numerals.

Recently, Convolutional Neural Network (CNN) has reported great success in character recognition task . A convolutional neural network has been widely used for classification and recognition of almost all the languages that have been reviewed for this systematic literature review .

### B. Kernel Methods

A number of powerful kernel-based learning models, e.g. Support Vector Machines (SVMs), Kernel Fisher Discriminant Analysis (KFDA) and Kernel Principal Component Analysis (KPCA) have shown practical relevance for classification problems. For instance, in the context of optical pattern, text categorization, time-series prediction, these models have significant relevance.

In support vector machine, kernel performs mapping of feature vectors into a higher dimensional feature space in order to find a hyperplane, which is linearly separates classes by as much margin as possible. Given a training set of labeled examples {(xi,yi),i=1…l} where xi∈ Rn and yi∈ {−1, 1}, a new test example x is classified by the following function:

f(x)=sgn(∑i=11αiyiK(xi,x)+b)(1)

1. K (*.,.)* is a kernel function
2. b is the threshold parameter of the hyperplane
3. αi are Lagrange multipliers of a dual optimization problem that describe the separating hyperplane

Before the popularization of deep learning methodology, SVM was one of the most robust technique for handwritten digit recognition, image classification, face detection, object detection, and text classification [75]. Kernel Fisher Discriminant Analysis (KFDA) and Kernel Principal Component Analysis (KPCA) are also some of the most significant kernel methods being used in offline handwritten character recognition system [76]. A number of researchers still believe that SVM performs better than most of the other techniques in classifying the handwritten characters. This is the reason why SVM is still being used for the purpose of classification of characters in HCR .

Previously, Boukharouba and Bennia  and Yang *et al.*  used SVM for recognition of Urdu and Arabic handwritten digits. SVMs have also been successfully implement in image classification and affect recognition , text classification  and face and object detection .

### C. Statistical Methods

Statistical classifiers can be parametric and non-parametric. Parametric classifiers have fixed (finite) number of parameters, and their complexity is not a function of the size of input data. Parametric classifiers are generally fast in learning concept and can even work with the small training set. Example of parametric classifiers is Logistic Regression (LR), Linear Discriminant Analysis (LDA), Hidden Markov Model (HMM) etc.

On the other hand, non-parametric classifiers are more flexible in learning concepts but usually grow in complexity with the size of input data. K Nearest Neighbor (K NN), Decision Trees (DT) are examples of non-parametric techniques as their number of parameters grows with the size of the training set.

#### 1) Non-Parametric Statistical Methods

One of the most used and easy to train statistical model for classification is k nearest neighbor (k NN) [42], [87], [88]. It is a non-parametric statistical method, which is widely used in optical character recognition. Non-parametric recognition does not involve a-priori information about the data.

k NN finds a number of training samples closest to a new example-based on target function. Based on the value of the targeted function, it infers the value of the output class. The probability of an unknown sample q belonging to class y can be calculated as follows:

p(y|q)=Wk=∑k∈KWk.1(ky=y)∑k∈KWk1d(k,q)(2)(3)

1. K is the set of nearest neighbors
2. ky the class of k
3. d (*k,* q ) the Euclidean distance of k from q , respectively.

Researchers have been found to use k NN for over a decade now, and they believe that this algorithm achieves relatively good performance for character recognition in their experiments performed on different datasets [2], [18], [62], [88].

k NN classifies object/ROI based on the majority vote of its neighbours (class) as it assigns class most prevalent among its k nearest neighbours. If k = 1, then the object is simply assigned to a class of that single nearest neighbour [57].

#### 2) Parametric Statistical Methods

As mentioned above, parametric techniques models concepts using fixed (finite) number of parameters as they assume sample population/training data can be modelled by a probability distribution that has a fixed set of parameters. In OCR research studies, generally, characters are classified according to some decision rules such as a maximum likelihood or Bayes method once parameters of the model are learned [36].

Hidden Markov Model (HMM) was one of the most frequently used parametric statistical method earlier in 2000.

HMM, models system/data that is assumed to be a Markov process with hidden states, wherein Markov process probability of one states only depends on previous state [36]. It was first used in speech recognition during the 1990s before researchers started using it in recognition of optical characters [89]–[90][91]. It is believed that HMM provides better results even when the availability of lexicons is limited [41].

### D. Template Matching Techniques

As the names suggest, template matching is an approach in which images (a small part of an image) is matched with a certain predefined template. Usually, template matching techniques employ a sliding window approach in which template image or feature are sliders on the image to determine the similarity between the two. Based on used similarity (or distance) metric classification of different objects are obtained [92].

In OCR, template matching technique is used to classify character after matching it with the predefined template(s) [93]. In literature, different distance (similarity) metrics are used, most common ones are Euclidean distance, city block distance, cross-correlation, normalized correlation etc.

In template matching, either template matching technique employs a rigid shape matching algorithm or deformable shape matching algorithm. Thus, creating a different family of template matching. Taxonomy of template matching techniques is presented in Figure 9.

One of the most applicable approaches for character recognition is deformable template matching (refer Figure 10) as different writers can write character by deforming them in a particular way specific to writer. In this approach, a deformed image is used to compare it with a database of known images. Thus, matching/classification is performed with deformed shapes as a specific writer could have deformed character in a particular way [36]. Deformable template matching is further divided into parametric and free form matching. Prototype matching, which is sub-class of parametric deformable matching, matching of done based on a stored prototype (deformed) [94].

Apart from the deformable template matching approach, second sub-class of template matching is rigid template matching. As the name suggests, rigid template matching does not take into account shape deformations. This approach usually works with features extraction/matching of the image with a template. One of the most common approaches used in OCR to extract shape features is Hough transform, like Arabic [95] and Chinese [96].

Second sub-class of rigid template matching is correlation- based matching. In this technique, initially, image similarity is calculated and based on similarity features from specific regions are extracted and compared [36], [97].

### E. Structural Pattern Recognition

Another classification technique that was used by OCR research community before the popularization of kernel methods and neural networks/deep learning approach was structural pattern recognition. Structural pattern recognition aims to classify objects based on a relationship between its pattern structures and usually structures are extracted using pattern primitives (refer Figure 11 for an example of pattern primitives), i.e. edge, contours, connected component geometry etc. One of such image primitive that has been used in OCR is Chain Code Histogram (CCH) [98], [99]. CCH effectively describes image/character boundary/curve, thus helping in classify character [57], [75]. Prerequisite condition to apply CCH for OCR is that image should be in binary format, and boundaries should be well defined. Generally, for handwritten character recognition, this condition makes CCH difficult to use. Thus, different research studies and publicly available datasets use/provide binarized images [87].

In research studies of OCR, structural models can be further subdivided on the basis of the context of structure, i.e. graphical methods and grammar-based methods. Both of these models are presented in the next two sub-sections.

#### 1) Graphical Methods

A graph (G ) is a way to mathematically describe a relation between connected objects and is represented by an ordered pair of nodes (N ) and edges (E ). Generally, for OCR, E represents the arc of writing stroke connecting N . The particular arrangement of N and E define characters/digits/alphabets. Trees (undirected graph, where the direction of the connection is not defined), directed graphs (where the direction of edge to a node is well defined) are used in different research studies to represent characters mathematically [101], [102].

As mentioned above, writing structural components are extracted using pattern primitives, i.e. edge, contours, connected component geometry etc. The relation between these structures can be defined mathematically using graphs (refer Figure 11 for an example showing how letter “R” and “E” can be modelled using graph theory). Then considering specific graph architecture different structures can be classified using graph similarity measure i.e. similarity flooding algorithm [103], SimRank algorithm [104], Graph similarity scoring [105] and vertex similarity method [106]. In one study [107], graph distance is used to segment overlapping and joined characters as well.

#### 2) Grammar Based Methods

In graph theory, syntactic analysis is also used to find similarities in structural graph primitives using the concept of grammar [108]. The benefit of using grammar concepts in finding the similarity in graphs comes from the fact that this area is well researched and techniques are well developed. There are different types of grammar based on restriction rules, for example, unrestricted grammar, context-free grammar, context-sensitive grammar and regular grammar. Explanation of these grammar and corresponding applied restrictions are out the scope of this survey article.

In OCR literature, usually, strings and trees are used to represent models based on grammar. With well-defined grammar, a string is produced that then can be robustly classified to recognize the character. The tree structure can also model hierarchical relations between structural primitives [92]. Trees can also be classified by analyzing grammar that defines the tree, thus classifying specific character [109].

**SECTION VI.**

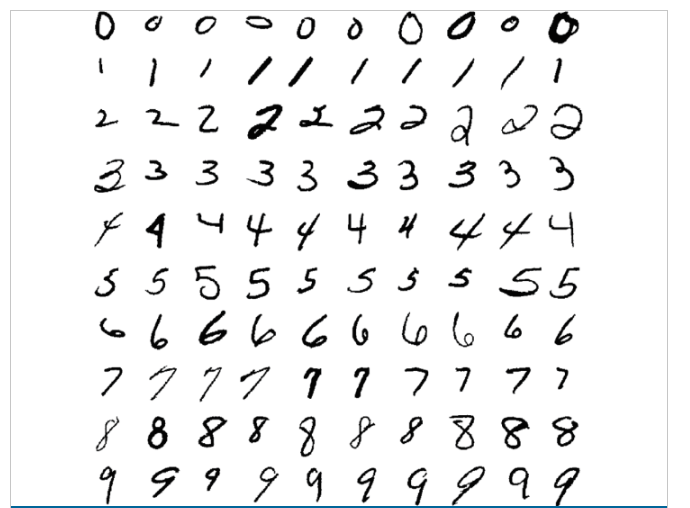
## Datasets

Generally, for evaluating and benchmarking different OCR algorithms, standardized databases are needed/used to enable a meaningful comparison [55]. Availability of a dataset containing enough amount of data for training and testing purpose is always a fundamental requirement for a quality research [110], [111]. Research in the domain of optical character recognition mainly revolves around six different languages, namely, English, Arabic, Indian, Chinese, Urdu and Persian/Farsi script. Thus, there are publicly available datasets for these languages such as MNIST, CEDAR, CENPARMI, PE92, UCOM, HCL2000 etc.

Following subsections presents an overview of most used datasets for the above mentioned languages.

### A. CEDAR

This legacy dataset, CEDAR, was developed by the researchers at the University of Buffalo in 2002 and is considered among the first few large databases of handwritten characters [40]. In CEDAR, the images were scanned at 300 dpi. Example character images from the CEDAR database are shown in Figure 12.

**FIGURE 12.**

**FIGURE 13.**

**FIGURE 14.**

### B. Chars74k

Chars74k [112] dataset was introduced by researchers at the university of surrey in 2009. The dataset contains 74,000 images of English and Kannada (Indian) scripts. The database contains street scenes taken in Bangalore, India. One thousand nine hundred twenty-two images of signboards, hoardings, advertisements and products in supermarkets were photographed. Segmentation of individual characters was done manually, and results were presented in bounding box segmentation. Bag of visual words technique was used for object categorization, and eventually, 62 different classes were created for English and 657 classes for Kannada.

A number of researchers have used CHARS74k dataset for recognition of Kannada script. Naiemi [78] applied histogram of oriented gradients features on CHARS74k dataset for spam image detection, while [113] used the dataset for recognizing characters in early Indian printed documents. Joe *et al.* [114] used CNN to recognize offline handwritten characters written in Kannada script.

It is to be noted that Kannada is one of many Indian scripts we have included in this research. There are various datasets for Indian language, depending on the script that has been used. For example, CMATERDB is a dataset for Indian script called Bangla [115], [116] and Kaggle’s Tamil handwritten character dataset is another such dataset for Tamil script [117].

### C. MNIST

The MNIST dataset is considered as one of the most used/ cited dataset for handwritten digits [30], [42], [118]–[119][120][121]. It is the subset of the NIST dataset, and that is why it is called modified NIST or MNIST. The dataset consists of 60,000 training and 10,000 test images. Samples are normalized into 20×20 grayscale images with reserved aspect ratio, and the normalized images are of size 28×28 . The dataset greatly reduces the time required for pre-processing and formatting, because it is already in a normalized form.

### D. UCOM

The UCOM is an Urdu language dataset available for research [122]. The authors claim that this dataset could be used for both character recognition as well as writer identification. The dataset consists of 53,248 characters and 62,000 words written in nasta’liq (calligraphy) style, scanned at 300 dpi. The dataset was created based on the writing of 100 different writers where each writer wrote 6 pages of A4 size. The dataset evaluation is based on 50 text line images as train dataset and 20 text line images as test dataset with reported error rate between 0.004 -0.006%. Example characters from the dataset are presented in Figure 15.

### E. IFN/ENIT

The IFN/ENIT [37] is the most popular Arabic database of handwritten text. It was developed in 2002 by the researchers at Technical University Braunschweig, Germany for the advancement of research and development of Arabic handwriting recognition systems. The dataset contains 26459 handwritten images of the names of towns and villages in Tunisia. These images consist of 212,211 characters written by 411 different writers, refer Figure 16. Since its inception, the dataset has been widely used by the researchers for the efficient recognition of Arabic characters .

### F. CENPARMI

The CENter for Pattern Recognition and Machine Intelligence (CENPARMI) introduced the first version of Farsi dataset in 2006 [51], [125]. This dataset contains 18,000 samples of Farsi numerals. These numerals are divided into 11,000 training, 2,000 verification and 5,000 samples for testing purpose.

Another similar, but larger dataset of Farsi numerals was produced by Khosravi and Kabir [51] in 2007. This dataset contains 102,352 digits extracted from registration forms of high school and undergraduate students. Later in 2009 [126], CENPARMI released another larger, extended version of Farsi dataset. This larger dataset contains 432,357 images of dates, words, isolated letters, isolated digits, numeral strings, special symbols, and documents. Refer Figure 17 for examples images from CENPARMI Farsi language dataset.

### G. HCL2000

The HCL2000 is a handwritten Chinese character database, refer Figure 18 to see sample images. The dataset is publicly available for researchers. The dataset contains 3,755 frequently used Chinese characters written by 1,000 different subjects. The database is unique in a way that it contains two sub-datasets, one is handwritten Chinese characters dataset, while the other is corresponding writer’s information dataset. This information is provided so that research can be conducted not only based on character recognition, but also on the writer’s background such as age, gender, occupation and education .

### H. IAM

The IAM [128] is a handwritten database of English language based on Lancaster-Oslo/Bergen (LOB) corpus. Data were collected from 400 different writers who produced 1,066 forms of English text containing a vocabulary of 82,227 words. Data consists of full English language sentences. The dataset was also used for writer identification [48]. Researchers were able to successfully identify writer 98% of the time during experiments on IAM dataset. Writing sample from the IAM dataset are presented in Figure 19.

**SECTION VII**

## Languages

As mentioned above, researchers working in the domain of optical character recognition have mainly investigated six different languages, which are English, Arabic, Indian, Chinese, Urdu and Persian. This is one of the future work to built OCR systems for other languages as well.

According to the United Nations Educational, Scientific and Cultural Organization (UNESCO) report on “world’s languages in danger,” at least 43% of languages spoken in the world are endangered [129]. These large number of languages need the attention of OCR research community as well to preserve this heritage from extinction or at least to build such a system that translates documents from endangered languages to electronic form for reference. Data from UNESCO’s report on “world’s languages in danger” is presented in Figure 20.

This section presents state-of-the-art results for six languages which are usually studied by researchers.

### A. English Language

The English Language is the most widely used language in the world. It is the official language of 53 countries and articulated as a first language by around 400 million people. Bilinguals use English as an international language. Character recognition for the English language has been extensively studied throughout many years. In this systematic literature review, the English language has the highest number of publications, i.e. 45 publications after concluding the study selection process (refer Section II-D and Section III-D). The OCR systems for the English language occupy a significant place as a large number of studies have been done in the era of 2000–2018 on the English language.

The English language OCR systems have been used successfully in a wide array of commercial applications. The most cited study for English language handwritten OCR is by Plamondon and Srihari [35] in 2000, which have more than 2900 citations, refer Table 3. The objective of the research by Plamondon *et al.* was to present a broad review of state of the art in the field of automatic processing of handwriting. This paper explained the phenomenon of pen-based computers and achieved the goal of automatic processing of electronic ink by mimicking and extending the pen-paper metaphor. To identify the shape of the character, structural and rule-based models like (SOFM) self-organized feature map, (TDNN) time-delay neural network and (HMM) hidden Markov model was used.

Another comprehensive overview of character recognition presented in [36] by Arica *et al.* has more than 500 citations. Arica *et al.* concluded that characters are natural entities, and it is practically impossible for character recognition to impose a strict mathematical rule on the patterns of characters. Neither the structural nor the statistical models can signify a complex pattern alone. The statistical and structural information for many characters pattern can be combined by neural networks (NNs) or harmonic markov models (HMM).

Connell and Jain [9] demonstrated a template-based system for online character recognition, which is capable of representing different handwriting styles of a particular character. They used decision trees for efficient classification of characters and achieved 86% accuracy.

Every language has specific way of writing and have some diverse features that distinguished it with other language. We believe that to efficiently recognize handwritten and machine printed text of the English language, researchers have used almost all of the available feature extraction and classification techniques. These feature extraction and classification techniques include but not limited to HOG [130], bidirectional LSTM [131], directional features [132], multilayer perceptron (MLP) , hidden markov model(HMM) , Artificial neural network (ANN) and support vector machine (SVM).

Recently trend is shifting away from using handcrafted features and moving towards deep neural networks. Convolutional Neural Network (CNN) architecture, a class of deep neural networks, has achieved classification results that exceed state-of-the-art results specifically for visual stimuli/input [138]. LeCun [20] proposed CNN architecture based on multiple stages where each stage is further based on multiple layers. Each stage uses feature maps, which are basically arrays containing pixels. These pixels are fed as input to multiple hidden layers for feature extraction and a connected layer, which detects and classifies object [55]. A recent study by [69] used fully convolutional neural network(FCNN) on IAM and RIMES datasets. Results were promising, and researchers achieved the character error rate(CER) and word error rate(WER) of 4.7%, 8.22%, 2.46%, 5.68% respectively. Jayasundara [139] proposed a novel technique called capsule networks(CapsNet) for the handwritten character recognition with very small datasets. Research claims that these techniques require a very small number of training samples for each class. These samples can be as low as 200. It is also claimed that the proposed technique can produce results similar to state-of-the-art systems, with only 10% of the data. When the proposed technique was applied to small datasets, it achieved the accuracy of 90.46%.

### B. Farsi/Persian Script

Farsi, also known as the Persian Language, is mainly spoken in Iran and partly in Afghanistan, Iraq, Tajikistan and Uzbekistan by approximately 120 million people. The Persian script is considered to be similar to Arabic, Urdu, Pashto and Dari languages. Its nature is also cursive, so the appearance of the letter changes with respect to positions. The script comprises of 32 characters, and unlike the Arabic language, the writing direction of the Farsi language is mostly but not exclusively from right to left.

Mozaffari *et al.* [140] proposed a novel handwritten character recognition method for isolated alphabets and digits of Farsi and Arabic language by using fractal codes. On the basis of the similarities of the characters, they categorized the 32 Farsi alphabets into 8 different classes. A multilayer perceptron (MLP) (refer Figure 8 for an overview of MLP) was used as a classifier for this purpose. The classification rate for characters and digits were 87.26% and 91.37% respectively.

However, in another research [141], researchers achieved a recognition rate of 99.5% by using RBF kernel-based support vector machine. Broumandnia and Shanbehzadeh [142] conducted research on Farsi character recognition and claims to propose the fastest approach of recognizing Farsi character using Fast Zernike wavelet moments and artificial neural networks (ANN). This model improves on average recognition speed by 8 times.

Liu and Suen [66] presented results of handwritten Bangla and Farsi numeral recognition on binary and grayscale images. The researchers applied various character recognition methods and classifiers on the three public datasets such as ISI Bangla numerals, CENPARMI Farsi numerals, and IFHCDB Farsi numerals and claimed to have achieved the highest accuracies on the three datasets, i.e. 99.40%, 99.16%, and 99.73%, respectively.

In another research, Boukharouba and Bennia [75] proposed SVM based system for efficient recognition of handwritten digits. Two feature extraction techniques, namely, chain code histogram (CCH) [143] and white-black transition information, were discussed. The feature extraction algorithm used in the research did not require digits to be normalized. SVM classifier, along with RBF kernel method, was used for classification of handwritten Farsi digits named ‘hoda’. This system maintains high performance with less computational complexity as compared to previous systems as the features used were computationally simple.

Researchers have also used Convolutional Neural Network (CNN) in conjunction with other techniques for the recognition of characters. These techniques have been applied on different datasets to check the accuracy of techniques [74], [87], [144]–[145][146].

### C. Urdu Language

Urdu is curvasive language like Arabic, Farsi and many other [147]. In the Urdu language, a notable early attempt to improve the methods for OCR is by Javed *et al.* in 2009 [148]. Their study focuses on the Nasta’liq (calligraphy) style-specific pre-processing stage in order to overcome the challenges posed by the Nasta’liq style of Urdu handwriting. The steps proposed include page segmentation into lines and further line segmentation into sub-ligatures, followed by base identification and base-mark association. 94% of the ligatures were accurately separated with proper mark association.

Later in 2009, the first known dataset for Urdu handwriting recognition was developed at Centre for Pattern Recognition and Machine Intelligence (CENPARMI) [149]. Sagheer *et al.* [149] focused on the methods involving data collection, data extraction and pre-processing. The dataset stores dates, isolated digits, numerical strings, isolated letters, special symbols and 57 words. As an experiment, Support Vector Machine (SVM) using a Radial Base Function/kernel (RBF) was used for classification of isolated Urdu digits. The experiment resulted in a high recognition rate of 98.61%.

To facilitate multilingual OCR, Hangarge and Dhandra [118] proposed a texture-based method for handwritten script identification of three major scripts: English, Devnagari and Urdu. Data from the documents were segmented into text blocks and/or lines. In order to discriminate the scripts, the proposed algorithm extracts fine textural primitives from the input image based on stroke density and pixel density. For experiments, k -nearest neighbour classifier was used for classification of the handwritten scripts. The overall accuracy for tri-script and bi-script classification peaked up to 88.6% and 97.5% respectively.

A study by Pathan *et al.* [7] in 2012 proposed an approach based on the invariant moment technique to recognize the handwritten isolated Urdu characters. A dataset comprising of 36800 isolated single and multi-component characters was created. For multi-component letters, primary and secondary components were separated, and invariant moments were calculated for each. The researchers used SVM for classification, which resulted in an overall performance rate of 93.59%. Similarly, Raza *et al.* [150] created an offline sentence database with automatic line segmentation. It comprises of 400 digitised forms by 200 different writers.

Obaidullah *et al.* [151] proposed a handwritten numeral script identification (HNSI) framework to identify numeral text written in Bangla, Devanagari, Roman and Urdu. The framework is based on a combination of daubechies wavelet decomposition [152] and spatial domain features. A dataset of 4000 handwritten numeral word image for these scripts was created for this purpose. In terms of average accuracy rate, multi-layer perceptron (MLP) (refer Figure 8 for a pictorial depiction of MLP) proves to be better than NBTree, PART, Random Forest, SMO and Simple Logistic classifiers.

In 2018, Asma and Kashif [153] presented a comparative analysis of raw images and meta-features from UCOM dataset. CNN (Convolutional Neural Network) and an LSTM (Long short-term memory), which is a recurrent neural network-based architecture were used on Urdu language dataset. Researchers claim that CNN provided accuracy of 97.63% and 94.82% on thickness graph and raw images, respectively. While the accuracy of LSTM was 98.53% and 99.33%. Naseer and Zafar [153] and Tayyab *et al.* [154] proposed an OCR model based on CNN and BDLSTM (Bi-Directional LSTM). This model was applied to a dataset containing Urdu news tickers, and results were compared with google’s vision cloud OCR. Researchers found that their proposed model worked better than google’s cloud vision OCR in 2 of the 4 experiments.

In 2019 Ahmed *et al.* [155] proposed a technique based on one-dimensional BLSTM classifier that used recurrent neural network(RNN), long-short term memory(LSTM) and bidirectional recurrent neural networks(BRNN) for the recognition of handwritten Urdu written in Nasta’liq style. Researchers also presented a new dataset of 500 writers named Urdu-Nasta’liq handwritten dataset (UNHD). Researchers claim to have achieved very good accuracy in recognizing the characters. The error rate was 6.04–7.93% during various experiments. During another study, Rafeeq *et al.* [156] used a deep neural network with dropout regularization. Ligatures were categorized, and the K-Means algorithm is used to cluster the ligatures. Researchers claim that their proposed technique achieved 94.71% accuracy as compared to neural networks which achieved only 74.31% accuracy.

### D. Chinese Language

Our research includes 23 research publications on the OCR system of Chinese language after concluding the study selection process (refer Section II-D and Section III-D). One of the Earliest research on the Chinese language was done in 2000 by Fu *et al.* [157]. The researchers used self-growing probabilistic decision-based neural networks (SPDNNs) to develop a user adaptation module for character recognition and personal adaption. The resulting recognition accuracy peaked up to 90.2% in ten adapting cycles.

Later in 2005, a comparative study of applying feature vector-based classification methods to character recognition by Liu and Fujisawa [67] found that discriminative classifiers such as an artificial neural network (ANN) and support vector machines (SVM) gave higher classification accuracies than statistical classifiers when the sample size was large. However, in the study SVM demonstrated better accuracies than neural networks in many experiments.

In another study Bai and Huo [45] evaluated the use of 8-directional features to recognize online handwritten Chinese characters. Following a series of processing steps, blurred directional features were extracted at uniformly sampled locations using a derived filter, which forms a 512-dimensional vector of raw features. This, in comparison to an earlier approach of using 4-directional features, resulted in a much better performance.

In 2009, Zhang *et al.* [127] presented HCL2000, a large-scale handwritten Chinese Character database. It stores 3,755 frequently used characters along with the information of its 1000 different writers. HCL2000 was evaluated using three different algorithms; Linear Discriminant Analysis (LDA), Locality Preserving Projection (LPP) and Marginal Fisher Analysis (MFA). Prior to the analysis, the Nearest Neighbor classifier assigns input image to a character group. The experimental results show MFA and LPP to be better than LDA.

Yin *et al.* [53] proposed ICDAR 2013 competition which received 27 systems for 5 tasks – classification on extracted feature data, online/offline isolated character recognition and online/offline handwritten text recognition. Techniques used in the systems were inclusive of LDA, Modified quadratic discriminant function (MFQD), Compound Mahalanobis Function (CMF), convolutional neural network (CNN) and multilayer perceptron (MLP). It was explored that the methods based on neural networks proved to be better for recognizing both isolated character and handwritten text.

During the study in 2016 on accurate recognition of multilingual scene characters, Tian *et al.* [130] proposed an extension of Histogram of Oriented Gradient (HOG), Cooccurrence HOG (Co-HOG) and Convolutional Co-HOG (ConvCo-HOG) features. The experimental results show the efficiency of the approaches used and higher recognition accuracy of multilingual scene texts.

In 2018, researchers on Chinese script used neural networks to recognize CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) recognition [158], Medical document recognition [159], License plate recognition [160] and text recognition in historical documents [161]. Researchers used Convolutional Neural Network(CNN) [158], [161], Convolutional Recurrent Neural Network(CRNN) [159] and Single Deep Neural Network(SDNN) [160] during these studies.

During 2019 [162], [163] used techniques based on recurrent neural network(RNN) for the recognition of online and offline handwritten text, respectively. On the other hand, Gan *et al.* [73] used 1-dimensional CNN for the recognition of online handwritten Chinese characters. 1-dimensional CNN seems to have performed better as recognition accuracy of [73] is 98.1% as compared to [163] where the accuracy of 83% was achieved. Zhu *et al.* [164] proposed a new neural network structure for Chinese handwritten character recognition. Researchers adaptively assigned different weights to category-classifiers depending on the quality of data. Maximum accuracy of 93.74% was achieved during the experiments on three different datasets.

### E. Arabic Script

Research on handwritten Arabic OCR systems has passed through various stages over the past two decades. Studies in the early 2000s focused mainly on the neural network methods for recognition and developed variants of databases [165]. In 2002, Pechwitz *et al.* [37] developed the first IFN/ENIT-database to allow for the training and testing of Arabic OCR systems. This is one of the highly cited databases and has been cited more than 470 times. Another database was developed by Mozaffari *et al.* [166] anMozaffari and Soltanizadeh [167] in 2006. It stores grey-scale images of isolated offline handwritten 17,740 Arabic/Farsi numerals and 52,380 characters. Another notable dataset containing Arabic handwritten text images was introduced by Mezghani *et al.* [168]. The dataset has an open vocabulary written by multiple writers (AHTID/ MW). It can be used for word and sentence recognition, and writer identification [169].

A survey by Lorigo and Govindaraju [18] provides a comprehensive review of the Arabic handwriting recognition methodologies and databases used until 2006. This includes research studies carried out on IFN/ENIT database. These studies mostly involved artificial neural networks (ANNs), Hidden Markov Models (HMM), holistic and segmentationbased recognition approaches. The limitations pointed out by the review included restrictive lexicons and restrictions on the text appearance.

In 2009, Graves and Schmidhuber [24] introduced a globally trained offline handwriting recognizer based on multi-directional recurrent neural networks and connectionist temporal classification. It takes raw pixel data as input. The system had an overall accuracy of 91.4%, which also won the international Arabic recognition competition.

Another notable attempt for Arabic OCR was made by Lutf *et al.* [170] in 2014, which primarily focused on the speciality of the Arabic writing system. The researcher proposed a novel method with minimum computation cost for Arabic font recognition based on diacritics. Flood-fill based and clustering-based algorithms were developed for diacritics segmentation. Further, diacritic validation is done to avoid misclassification with isolated letters. Compared to other approaches, this method is the fastest with an average recognition rate of 98.73% for 10 most popular Arabic fonts.

An Arabic handwriting synthesis system devised by Elarian *et al.* [171] in 2015 synthesizes words from segmented characters. It uses two concatenation models: ExtendedGlyphs connection and the Synthetic-Extensions connection. The impact of the results from this system shows significant improvement in the recognition performance of an HMM-based Arabic text recognizer.

Akram *et l.* [172] discussed an analytical approach to develop a recognition system based on HMM Toolkit (HTK). This approach requires no priori segmentation. Features of local densities and statistics are extracted using a vertical sliding windows technique, where each line image is transformed into a series of extracted feature vectors. HTK is used in the training phase, and Viterbi algorithm is used in the recognition phase. The system gave an accuracy of 80.26% for words with “Arabic-numbers” database and 78.95% with IFN/ENIT database.

In a study conducted in 2016 by Elleuch *et al.* [173], convolutional neural network (CNN) based on support vector machine (SVM) is explored for recognizing offline handwritten Arabic. The model automatically extracts features from raw input and performs classification.

In 2018, researchers applied the technique of DCNN (deep CNN) for recognizing the offline and handwritten Arabic characters [174]. An accuracy of 98.86% was achieved when the strategy of DCNN using transfer learning was applied to two datasets. In another similar study [175] an OCR technique based on HOG (Histograms of Oriented Gradient) [176] for feature extraction and SVM for character classification was used on the handwritten dataset. The dataset contained names of Jordanian cities, towns and villages yielded an accuracy of 99%. However, when the researchers used multichannel neural network for segmentation and CNN for recognition of machine-printed characters, the experiments on 18pt font showed an overall accuracy of 94.38%.

In 2019, Sahlol *et al.* [177] applied hybrid machine learning approach on CENPARMI dataset. The technique used the rough neighbourhood sets along with binary whale optimization algorithm. Researcher claims that the proposed technique consumes less amount of time in recognizing the characters as compared to VGGnet, Resnet, Nasnet, Mobilenet, Inception, and Xception. Alrehali *et al.* [71] used CNN on various datasets of historical Arabic manuscripts and achieved an accuracy of 74% to 88%. In an interesting study Ali and Suresha [79] used classifier fusion technique based on a fusion of features moments invariants(MI), runlength matrix(RLM), statistical properties of intensity histogram(SFIH) and wavelet decomposition(WD) and classifiers modified quadratic discriminate functions(MQDF), support vector machine(SVM) and random forest(RF). Researcher claim that the fusion technique provided accuracy of 97% to 99.8%, which is among the highest in Arabic handwritten character recognition.

### F. Indian Script

Indian script is collection of scripts used in the sub-continent namely Devanagari [138], Bangla [116], Hindi [178], Gurmukhi [63], Kannada [179] etc. One of the earliest research on Devanagari (Hindi) script was proposed in 2000 by Lehal and Bhatt [180]. The research was conducted on Devanagari script and English numerals. The researchers used data that was already in an isolated form in order to avoid the segmentation phase. The research is based on statistical and structural algorithms [181]. The results of Devanagari scripts were better than English numerals. Devanagari had a recognition rate of 89% with 4.5 confusion rate, while English numerals had a recognition rate of 78% with confusion rate of 18%.

Patil and Subbareddy [182] was the first researcher to use neural network approach for the identification of Indian documents. The researchers propose a system capable of reading English, Hindi and Kannada scripts. A modular neural network was used for script identification while a two-stage feature extraction system was developed, first to dilate the document image and second to find average pixel distribution in the resulting images.

Sharma *et al.* [46] proposed a scheme based on quadratic classifier for the recognition of the Devanagari script. The researchers used 64 directional features based on chain code histogram [143] for feature recognition. The proposed scheme resulted in 98.86% and 80.36% accuracy in recognizing Devanagari characters and numeral, respectively. Fivefold cross-validation was used for the computation of results.

Two research studies [50], [183] presented in 2007 were based on the use of fuzzy modelling for character recognition of Indian script. The researchers claim that the use of reinforcement learning on a small database of 3500 Hindi numerals helped achieve a recognition rate of 95%.

Another research carried out on Hindi numerals [25] used a relatively large dataset of 22,556 isolated numeral samples of Devanagari and 23,392 samples of Bangla scripts. The researchers used three Multi-layer perceptron classifiers to classify the characters. In case of a rejection, a 4th perceptron was used based on the output of the previous three perceptrons in a final attempt to recognize the input numeral. The proposed scheme provided 99.27% recognition accuracy vs the fuzzy modelling technique, which provided the accuracy of 95%.

Desai [28] used neural networks for the numeral recognition of Gujrati script. The researcher used a multi-layer feed-forward neural network for the classification of digits. However, the recognition rate was low at 82%.

Kumar *et al.* [184], [185] proposed a method for line segmentation of handwritten Devanagari text. An accuracy of 91.5% for line segmentation and 98.1% for word segmentation was achieved. Perwej and Chaturvedi [186] used backpropagation based neural network for the recognition of handwritten characters. The results showed that the highest recognition rate of 98.5% was achieved. Obaidullah *et al.* [151] proposed Handwritten Numeral Script Identification or HNSI framework based on four indic scripts, namely, Bangla, Devanagari, Roman and Urdu. The researchers used different classifiers, namely NBTree, PART, Random Forest, SMO, Simple Logistic and MLP and evaluated the performance against the true positive rate. Performance of MLP was found to be better than the rest. MLP was then used for bi and tri-script identification. Bi-script combination of Bangla and Urdu gave the highest accuracy rate of 90.9% on MLP, while the highest accuracy rate of 74% was achieved in a tri-script combination of Bangla, roman and Urdu.

In a multi dataset experiment [116], researchers applied a lightweight model based on 13 layers of CNN with 2-sub layers on four datasets of Bangla language. An accuracy of 98%, 96.81%, 95.71%, and 96.40% was achieved when the model was applied on CMATERdb, ISI, BanglaLekhaIsolated dataset and mixed datasets respectively. CNN based model was also applied on ancient documents written in Devanagari or Sanskrit script in another study. Results, when compared with Google’s vision, OCR gave an accuracy of 93.32% vs 92.90%.

In 2019 sahare and Dhok [77] conducted research on old Indian documents in Devanagari script. These documents had text diffusion due to degrading. Researchers used graph distance theory to carve up the characters that were overlapped due to diffusion before apply support vector machine to confirm the segmentation results. Paper also proposes a set of features based on the geometrical properties of the characters. These features claim to provide character recognition accuracy of 98.8% and 99.6%. During a study on Kannada, researchers used the HOG feature extraction method along with SVM classifier for spam image detection with an accuracy of 94.2% [78]. Research on Indian scripts is very diverse, and a number of researchers are involved in research on multiple scripts. This is the reason why a number of research articles on character recognition of Indian scripts are growing each year. researchers have used techniques like Tesseract OCR and google multilingual OCR [113], Convolutional Neural Network (CNN) [70], [114], Deep Belief Network with the distributed average of gradients feature [187],

Modified Neural Network with the aid of elephant herding optimization [188], VGG (Visual Geometry Group) [117] and SVM classifier with the polynomial and linear kernel [80].

**SECTION VIII.**

## Research Trends

Characters written by different individuals create large intraclass variability, which makes it difficult for classifiers to perform robustly. Lately, the research in the domain of optical character recognition has moved towards a deep learning approach [189], [190] with little emphasis on handcrafted features. Deep learning approach has produced improved classification accuracy at the cost of increased computational complexity, especially during the training phase.

In this section, we have analyzed hand character recognition research trend in the last three years (2017-2019). Our analysis is summarized in Table 5. Table 5 includes script under investigation, techniques or classification technique employed for OCR, year of publication and respective reference number. This table gives a holistic view of how researchers working on some of the widely used languages are trying to solve the problem of optical character recognition.

**TABLE 5**Summary of Frequently Used Feature Extraction and Classification Techniques: Data Corresponding to Last Three Years (2017-2019). Studies Corresponding to “Indian” Script do Include Research on Scripts Belonging to Devanagari, Bangla, Hindi, Kannada etc

Table 5 highlights the fact that the bulk of recent publications have employed a deep learning approach in some form. Especially CNN is being used extensively for the recognition of optical characters. This is partially due to the availability of large datasets. Researchers usually employ a deep learning approach for a language that has large enough dataset for deep learning to learn meaningful model. As stated above, although frameworks based on deep learning methods have obtained improved classification accuracy but at the cost of increased computational complexity. There are few recent studies that have utilized classical feature extraction approach in combination of feature selection algorithms and have obtained state-of-the-art result, for example

# METHODOLOGY

Expansion of software without planning and search is that the subject of future projects. we are able to edit the small print of the documents smoothly and that we can reuse structured details whererequired.

Automatic data entry with OCR is one in every of the foremost cost- effective, labor-intensive technology Recognition Recognitionoflatest font characters by program is incredibly easy and fast. What does the long run hold for OCR? they're adequately supplied by commercial manufacturers and are adequate research and development dollars, OCR may be a robust tool for data entry within the future applications. However, limited availability of money within the short term is feasible limit the expansion of this technology. However, if given the proper impetus and encouragement, a Many benefits are often provided through the OCR system. Of course: - Grid infrastructure utilized in the implementation of Optical Character Recognition The pace of coaching and awareness is greatly enhanced and increased in practice easy to use. Many programs are available where you will wish to read handwritten inputs. Reading handwriting may be a very difficult task considering the variations within the standardpen.

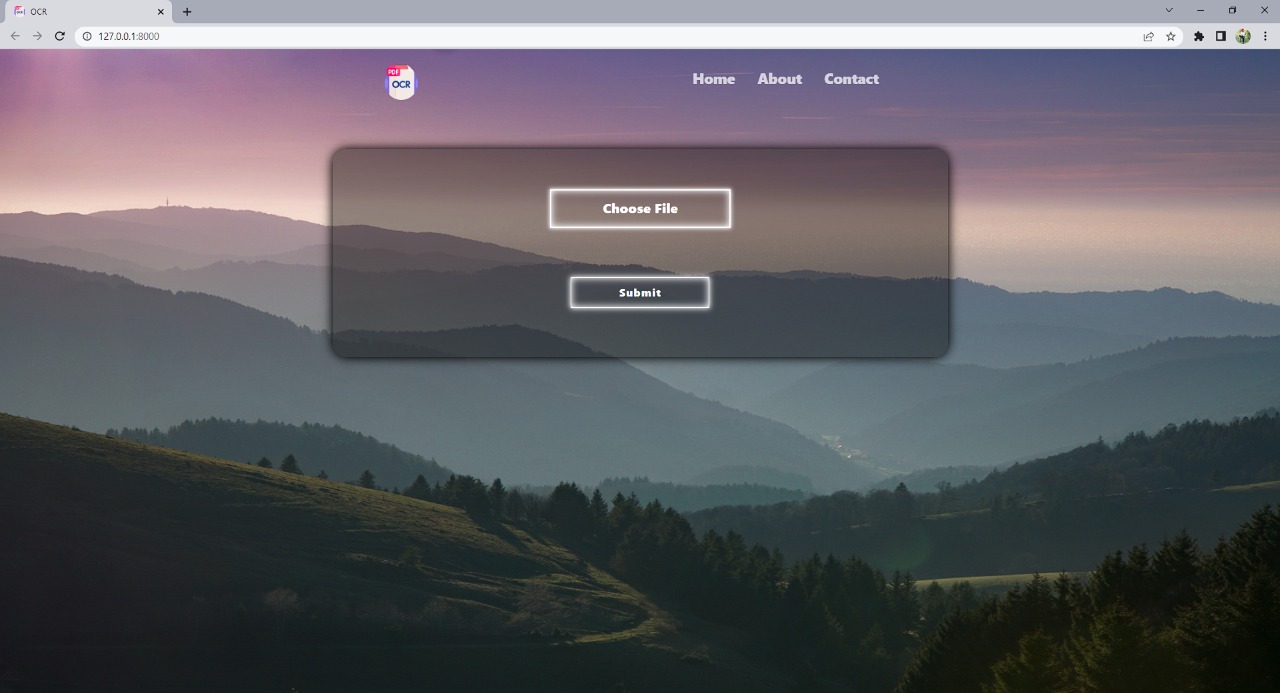
Nevertheless, progress is being mad system are often used effectively to hurry up the interpretation of image-based texts systematic texts that are now easy to search out, search, and process. FUTURE DEVELOPMENTS Optical Character Recognition software are often developed within the future with a spread of features of methods .

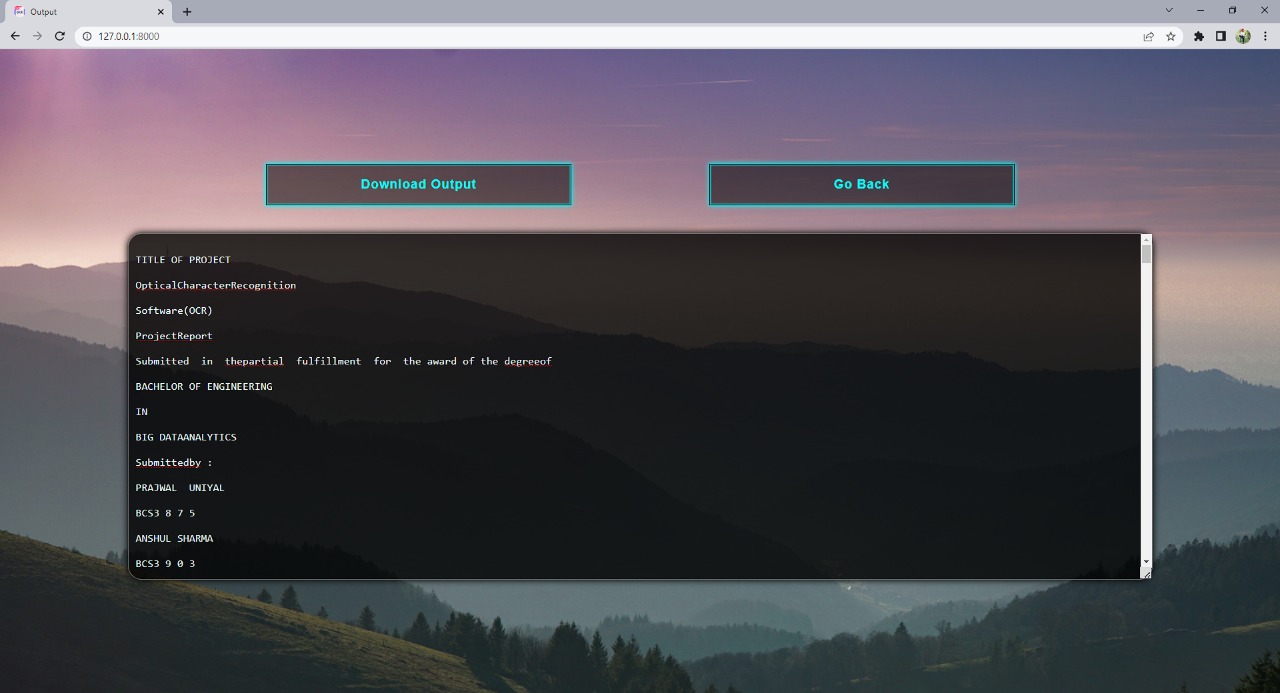
# RESULTS AND DISCUSSION

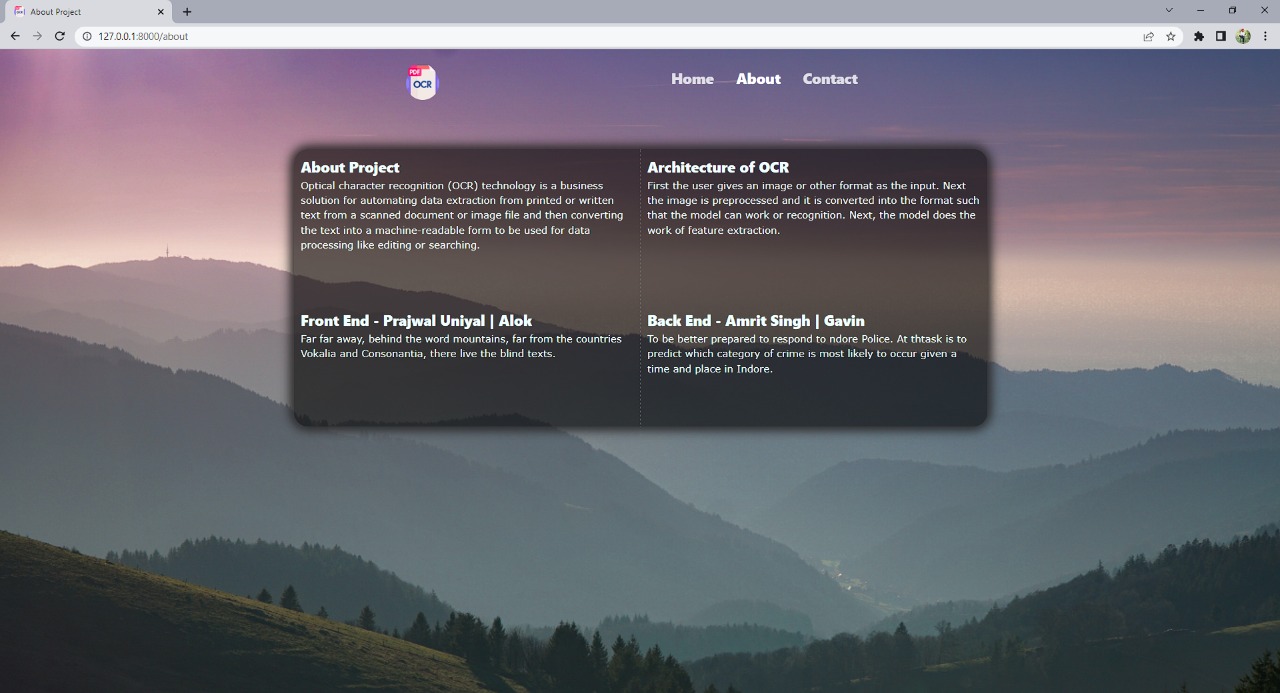
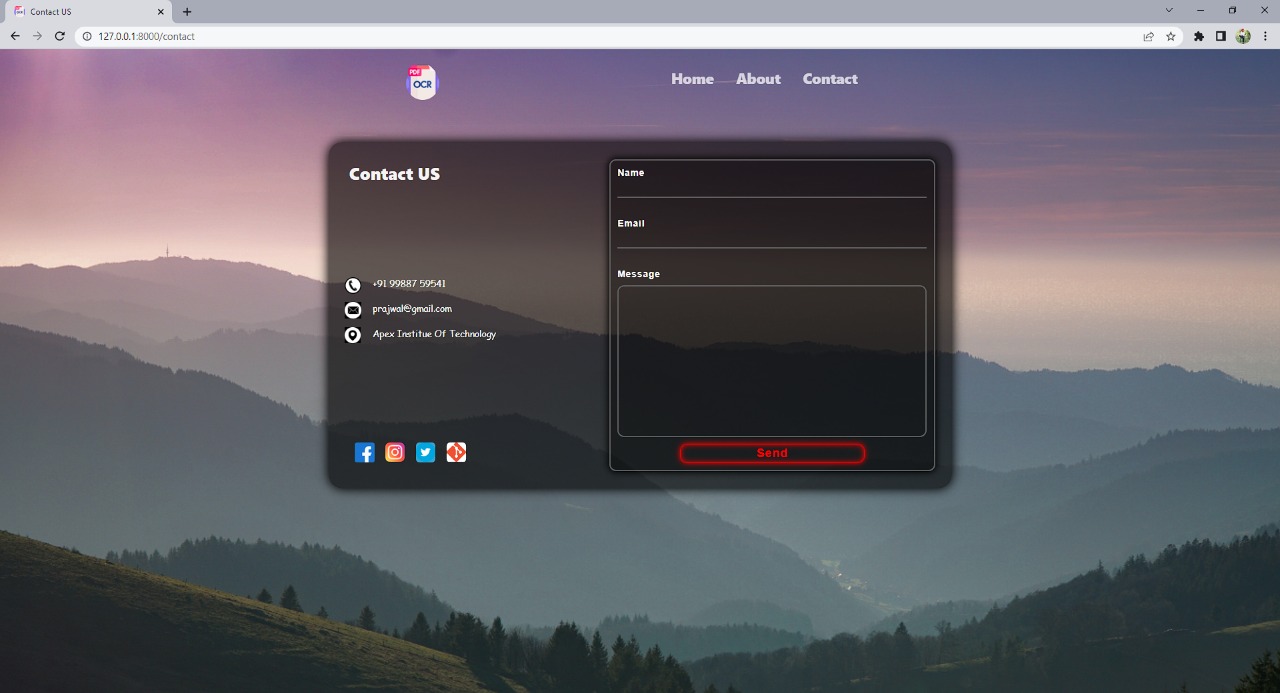
What does the longer term hold for OCR? Given enough entrepreneurial designers and sufficient research and development dollars, OCR can become a robust tool for future data entry applications. However, the limited availability of funds during a capital-short environment could restrict the expansion of this technology. But, given the right impetus and encouragement, a lotof advantages is provided by the OCR system. They are:- The automated entry of knowledge by OCR is one in every of the foremost attractive, labor reducing technology the popularity of recent font characters by the system is incredibly easy and quick. we are able to edit the knowledge of the documents more conveniently and that we can reuse the edited information as and whenrequired.

The extension to software aside from editing and searching istopic for future works. The Grid infrastructure utilized in the implementation of Optical Character Recognition system is efficiently accustomed speed up the interpretation of image based documents into structured documents that are currently easy toget,

search and process. FUTURE ENHANCEMENTS The Optical Character Recognition software may be enhanced within the future in numerous kinds of ways such as: Training and recognition speeds is increased greater and greater by making it more user-friendly. 86 Many applications exist where it'd be desirable to read handwritten entries. Reading handwriting may be a very difficult task considering the differences that exist in ordinary penmanship. However, progress is being made.







## Conclusion and Future Work

### A. Conclusion

1. Optical character recognition has been around for the last eight (8) decades. However, initially, products that recognize optical characters were mostly developed by large technology companies. Development of machine learning and deep learning has enabled individual researchers to develop algorithms and techniques, which can recognize handwritten manuscripts with greater accuracy.
2. In this literature review, we systematically extracted and analyzed research publications on six widely spoken languages. We explored that some techniques perform better on one script than on another, e.g. multilayer perceptron classifier gave better accuracy on Devanagri, and Bangla numerals [25], [140] but gave average results for other languages [119], [133], [134]. The difference may have been due to the fact of how specific technique models a different style of characters and quality of the dataset.
3. Most of the published research studies propose a solution for one language or even a subset of a language. Publicly available datasets also include stimuli that are aligned well with each other and fail to incorporate examples that correspond well with real-life scenarios,

i.e. writing styles, distorted strokes, variable character thickness and illumination [213].

1. It was also observed that researchers are increasingly using Convolutional Neural Networks(CNN) for the recognition of handwritten and machine-printed characters. This is due to the fact that CNN based architectures are well suited for recognition tasks where input is an image. CNN was initially used for object recognition tasks in images, e.g. the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 216]. AlexNet [215], GoogLeNet [216] and ResNet [217] are some of the CNN based architectures widely used for visual recognition tasks.

### B. Future Work

1. As mentioned in Section VII, research in OCR domain is usually done on some of the most widely spoken languages. This is partially due to non-availability of datasets on other languages. One of the future research direction is to conduct research on languages other than widely spoken languages, i.e. regional languages and endangered languages. This can help preserve the cultural heritage of vulnerable communities and will also create a positive impact on strengthening global synergy.
2. Another research problem that needs the attention of research community is to build systems that can recognize on-screen characters and text in different conditions in daily life scenarios, e.g. text in captions or news tickers, text on signboards, text on billboards etc. This is the domain of “recognition/classification/text in the wild”. This is a complex problem to solve as a system for such a scenario needs to deal with background clutters, variable illumination condition, variable camera angles, distorted characters and variable writing styles [213].
3. To build a robust system for “text in the wild”, researchers need to come up with challenging datasets that are comprehensive enough to incorporate all possible variations in characters. One such effort is [218]. In another attempt, the research community has launched “ICDAR 2019: Robustreading challenge on multilingual scene text detection and recognition” [219]. Aim of this challenge invites research studies that propose a robust system for multi-lingual text recognition in daily life or “in the wild” scenario. Recently report for this challenge has been published and winner methods for different tasks in the challenge are all based on different deep learning architectures, e.g. CNN, RNN or LSTM.
4. Characters written by different individuals create large intra-class variability, which makes it difficult for classifiers to perform robustly. Although with the increasing utilization of complex deep learning architectures, obtained classification accuracy has improved at the same time computational complexity (especially during the training phase of classifier) has grown. This creates a hurdle in the development of a real-time, robust system for hand character recognition.
5. Published research studies have proposed various systems for OCR but one aspect that needs to improve is the commercialization of research. Commercialization of research will help to build low-cost real-life systems for OCR that can turn lots of invaluable information into searchable/digital data [220].

# REFERENCES

1)<https://www.kaggle.com/datasets>

2)[https://www.flexmind.co/blog/how-to-use-ocr-in-power-automate-desktop/](https://www.flexmind.co/blog/how-to-use-ocr-in-power-automate-%20%20%20%20%20%20%20desktop/)

3)<https://www.techradar.com/in/best/best-ocr-software>4)<https://en.wikipedia.org/wiki/Optical_character_recognition>