The earliest form of writing known to human civilization is attributed to pictographic writing from Sumer in 3500 B.C. But, with passing of nearly 5600 years since the first pictographic writing, the human civilization has seen the evolution of languages numbering more than 100 across the globe with each of the scripts unique in their characters and scripts. The OCR developers, who attempted to preserve, the manuscripts as well as important writings, written in hundreds of languages over the ages for posterity faced a herculean task of converting them to electronic characters. A single OCR system, however efficiently developed, faced problem in its applicability as there are hundreds of languages with thousands of characters to recognize. This enormity of the study not only opened new frontiers for OCR developers with each language but also made it challenging and tricky with unique characters and vowels attributed to it.

But, early developments in computer, with convenience of preserving manuscripts as well as important documents, stored in comparatively small space, enhancing ease of transferability and convenience of multiple duplication, became a major inspiration to research for OCR developers. The advantages outweighed difficulties in recognizing thousands of characters in hundreds of languages and the scientists made all out effort to make a full proof OCR systems so that each and every recognized language, whether hand-written or printed, has an OCR system, which would recognize its characters and convert the same into electronic form with ease and with hundred percent accuracy.

The researcher attempts in this chapter to illustrate the historical development of OCR systems, process of converting text to electronic format, various methods used by different developers of this system as well as various products available in the market and development of OCR systems in Odia language by different researchers.

The main objective of Optical Character Recognition is to read the text, convert the text into electronic form so that the converted format could be saved into the electronic devices like computer, mobile, tablet etc. To accomplish the task of conversion, OCR
applications undergo many phases. These phases are technically named as Pre-processing, Segmentation, Feature Extraction, and Classification. Each phase applies to one of many different types of algorithms.

2.1. History of OCR

The history of OCR explains how it was developed throughout early research from the first OCR device to the current improved OCR applications. Interestingly OCR began as a hardware device. The problem at that time was finding a way to enable the computer to scan document images.

Even though with advent of modern day evolution of computer and allied hardware, innovation of a proper and efficient OCR system had to wait till 1928, when a patent was registered by Gustav Tauschek for basic OCR as a reading machine. The machine used light detecting photocells to recognize patterns in paper or card [5]. The predecessor to the patented version was created in 1870, which basically developed as an aid to visually handicap. The same was followed by an improved creation by the Russian scientist Tyurin in 1900. But as they fail to meet the expectations of OCR, the machines thus created, were being constantly improved upon.

Subsequent to evolution of photomultiplier machine in early 1950s, which was basically capturing the images of characters and text by mechanical and optical means of rotating disks and photomultipliers, the era of first real recorded OCR application ushered in. But real application of OCR had to wait till evolution of scanner machines [6].

With later development of flatbed scanner, the OCR got a new fillip, as the new flatbed machines were capable of high speed scanning as well as character recognition at higher speed and equally competent to scan a wide variety of forms and documents at a substantially reduced cost. Since then extensive researches have been carried out to further improve on the character recognition system thus achieved by new improved machines. Even though a large number of research studies have been published by various researchers in the area of character recognition, still an application of OCR with cent percent perfection is yet to be achieved.
As horizon of initial motive for developing OCR applications like processing sizeable amount of documents such as bank cheques, government records, credit card imprints and for mail sorting has expanded and also ever been expanding, the scope of OCR applications is also expanding.

2.2. Topics related to OCR

OCR development requires proper amalgamation of various techniques such as Digital Image Processing, Computer Vision, Pattern Recognition, Natural Language Processing, Artificial Neural Network and Optimization using Genetic approaches.

Processing of digital images by digital computers is referred as *Digital image processing* [7]. The digital image is composed of a fixed number of elements. Each element is a pixel on the image, that is identified by a location and value associated with it. Most of the steps of OCR handle the document image as a digital image and use different algorithms those are usually applied in image processing.

*Computer vision* attempts to duplicate the effort of human vision by electronically perceiving and understanding an image [8]. The machine extracts important information from the image which enable solving of some tasks in order to understand the image. The pre-processing stage employs computer vision algorithms.

*Pattern recognition* [9] is concerned with defined pattern domains (such as retina, fingerprint, voice and face) where it extracts some topographies from this pattern (such as geometrical or structural) and then applies a classification algorithm to identify the pattern based on previous known examples. The domain of OCR is a branch of pattern recognition.

*Natural language processing* (NLP) is a common research topic involving computer science and linguistics [10]. It is concerned with processing of human (natural) languages with the computers. This computer processing is generated using language engineering algorithms. OCR uses language engineering in generating and exploiting corpora.

*Artificial neural network* facilitates computers to carry out tasks, which are usually done by human brain [11]. Enabling computers to recognize text document with the help of
computational approaches for classification is considered as an artificial intelligence based approach.

2.3. Types of OCR System

Different approaches, which have been proposed by different researchers to design the OCR, are listed as follows:

- Rule based and Image based;
- Template and Feature Selection;
- Spatial domain and Transform domain;
- Offline and Online OCR System;
- Printed and Handwritten;

2.3.1. Rule based and Image based

In rule based approach of OCR systems, drawing styles of the images are captured as strict rules. The rules carry important information which are nothing but the piece of knowledge expressed by a human expert. Those are also called as If-Then or Production rules. It is the simplest method, but at the same time it is very tedious and accuracy of recognition is very less.

The image analysis techniques are used in a straightforward approach by image based OCR systems. At the initial stage of this, the system performs image enhancement functions, image segmentation process and finds some statistics of each of them. Based on these statistics, recognition can be achieved. The classical pattern recognition techniques are used by the image based approaches.

2.3.2. Template and Feature Selection

The image based OCR systems are further categorized into two ways:

- Template based and
- Feature based.
The template-based techniques are one of the most simplistic approaches. In this process, an unknown symbol is superimposed directly on the existing template pattern stored in databases and then degree of correlation between the two is used for the deciding about recognition. Early OCR systems employed only template approach. But this technique becomes ineffective in the presence of noise in the input document and also it gives poor result for Indian language as its character set are very similar as well as complex in nature. One of the biggest disadvantages of this technique is that it needs more space for storing of templates.

In case of feature extraction approach, for each symbol of the script, specific type of values, based on structural or geometrical properties, are generated. Those generated values are named as features of the corresponding symbols. Based on the features, the characters are classified. As the classification can be achieved by any machine learning approach that increases the accuracy of recognition, therefore the template based approaches are superimposed by the feature based approaches.

2.3.3. Spatial domain and Transform domain

The feature based approaches are further categorized into two types depending on the mode of extraction of features and are named as:

- Spatial domain and
- Transform domain.

The features of each character are derived directly from the pixel values of the image pattern in the spatial domain approaches. The pattern of the images is transformed into another space (i.e. frequency, time) and then features are derived, in the transformed domain approach.

Further, all of these classical pattern recognition techniques are dominated by neural network approaches.

2.4. Availability of OCR as a Product

Many commercial OCR applications or packages are available in the market. The main features that differentiate between them are as follows: degree of character
recognition accuracy, speed in conversion, support for different languages, support for page layout reconstruction and at last price to purchase. However products are mainly developed to be used with Roman Script.

Few of those commonly used OCR applications developed by different companies are listed as follows:

**ABBYY FineReader**

An international company ABBYY FineReader [12] first created omnifont OCR system by expanding its research and development effort during the year 1993 and onwards. The “ABBYY FineReader version 12” is the latest version of OCR produced by the company for many languages. Speed and accuracy of it virtually eliminate retyping. It is used by 20 million of people around the world for document processing and text recognition, but it did not satisfy to the needs of the language (i.e.Odia) concerned in this research.

**OmniPage Professional**

Nuance Company developed a latest OCR application “OmniPage Professional version 18” [13]. It recognizes text from scanned paper, pdf files and the digital camera images. It is the first OCR application that converts document images to original, even if having text with table, columns and graphics. This application is also having advance features so that it automatically corrects distortion if any in the document. It also does cater to the needs of various languages but not for the present language (i.e. Odia) under study.

**Readiris Pro**

Readiris Pro is manufactured by a large worldwide global company named I.R.I.S. started in 1987 and specialized in OCR applications. The other hardware optical character recognition products such as mobile scanner, pen scanner and digital pen are manufactured by this company. An OCR package “Readiris Pro 12” [14] was released for document recognition. Recently “Readiris Pro 15” is released only for Windows OS. Along with other features, it is also able to recognize hand written characters. It recognizes more than 120 different languages but still there is no application for Odia language.
**NovoVerus**

NovoDynamic, an In-Q-Tel company, developed a “NovoVerus” [15] OCR software for multicore architecture with cross platform facility. It is applicable for detecting and extracting the text from different languages. Also, it has the facility for enhancement of low quality and damaged photocopies of documents. It supports Roman, Asian as well as Middle Eastern languages but does not support Odia language.

**2.5. Literature Survey**

National Institute of Standards and Technology (NIST) [16] (founded in 1901 and now part of U.S. Department ) an Image Recognition Group has started working in large-scale OCR systems that are being used for Census and Internal Revenue Service (IRS) applications in the year 1989. NIST organized two OCR conferences to create awareness about OCR technology and its applications during the year 1992 and 1994. In these conferences, the applications of wide variety of pattern recognition methods to solve isolated character recognition has been discussed. The following subsections represent a detail effort of researchers for different languages and also for innovation of OCR system for Odia language too.

A substantial amount of research work on OCR systems for major languages of the world has been reported during last four decades and had been implemented successfully. Such languages include Roman, Latin, Arabic, Chinese, and Japanese scripts.

Kahan, et.al. [17] proposed an approach for Roman alphabets of any font and any size. A feature based method is adopted to develop OCR system by virtue of finding the strokes of characters using run length encoding technique. Clustering is performed thereafter with respect to the obtained shapes and Statistical Bayesian classifier is used for classification. It reports 99% of accuracy over a range of font size on a single font. Unfortunately the character set used in this research is not complex in shape as used by the Odia script. Also the rate of recognition is affected due to existence of dependency among certain considered features. The vectoring technique used for feature extraction and the classifier is slower, so time of execution can be improved.
Shrivastava et.al. [18] introduced 11 certain geometrical and topological representatives such as stroke, enclosures, curves etc. Those are calculated from pixels of English language. It considers 36 symbols and finds the 11 features for each. For training the system, it applies the Artificial Neural Network (ANN) classifier that uses back-propagation algorithm for learning. The recognition accuracy is 100% for the set of data trained upon, but the recognition rate decreases to 85.83% if the testing data are untrained data. This approach uses the templates matching technique instead of using any physical properties for recognition, which makes difficult for the complex font style Odia script. It also reports that for recognition of artistic font, the size of database needs to be increased.

Supriana, et. al. [19] subdivides the working procedure of OCR system for Arabic language into five stages named as pre-processing, segmentation, thinning, feature extraction, and classification. To eliminate the skew, it uses image moment, binarization that is achieved using median filter and Hilditch algorithm is used for thinning. They extracted 24 feature points considering the geometrical characteristics and applying tree classifier. Accuracy of achievement has been reported to be 99.9%. Because of the incompatibility of character segmentation and the dataset used for the training set of tree classifier, the performance of the entire system reaches to only 48.3 percentage.

Romero et.al. [20] uses the statistical measures, taken over the set of extracted features such as Stork width, Number of horizontal transition and Vertical transitions, Peripheral feature, Transitions from background to foreground pixels etc, from the Chinese characters. Probabilistic neural network is being used for classification purpose that allows computation of posterior probabilities on the classification result. It performs the testing, using 3964 characters and achieves the recognition rate of 97.04%.

2.5.1. Indic languages

India is a multi-lingual country and there are 22 languages recognized by the Indian constitution. Significant amount of research has been reported for various types of Indian scripts like, Hindi, Bangla, Devanagari, Tamil, Telugu, Malayalam, Urdu, Kannada, Gurmukhi, Gujarati etc. Indian Language Technology Proliferation and Development Centre (TDIL) [21] developed robust web OCR for printed Indian scripts such as Bangla, Devanagari, Tamil, Telugu, Malayalam, Gurmukhi, Kannada, Assamese and Urdu.
languages. Now days Centre for Development of Advance Computing (CDAC) [22] developed Bangla online form processing and online Bangla handwritten recognition system.

Pal et.al. [23] present a review of the work done about the Indian languages. They describe properties of alphabets used by different Indian scripts such as Hindi, Devnagari, Bangla, Gurumukhi, Urdu etc. and compare those with the English alphabet. They have discussed details of the traditional pattern recognition techniques i.e. Template and Feature based approaches used for character recognition. Apart from these, some modern techniques received the attention like Hidden Markov Model (HMM), Support Vector Machine (SVM), Fuzzy Rule, Mahalanobis and Hausdorff distance etc for character recognition.

Similarly Ghosh et. al. [24] emphasises on the importance of writing knowledge of the scripts before going into multi-script and multi-lingual environment of character recognition. Therefore, it represents different writing systems such as Logographic system, Syllabic system, Alphabetic system etc., used by different scripts. All Indian scripts belong to Brahmi family that follow the Alphabetic system. The survey report also concludes about the categorisation of methodologies developed so far into two groups: Structure based and Visual appearance based. For both of these groups, it discusses about the methods used for Page-wise identification, Paragraph block level recognition and Word or Character level identification. It concludes with a significant recognition accuracy achieved for the Indic languages such as Arabic, Urdu, Devnagari, Bengali, Gujarati and Tamil. Similarly, Nagy [25] published the detailed research work done in the field of character recognition for Indian language during last 20 years. Different techniques adopted by various researchers for the design of different stages of an OCR system are presented by Charlies et. al [26].

Kumar et.al. [27] introduced the challenges of designing an OCR for Indic script due to existence of large number of graphemes and their combination to form a new one. Three different semiautomatic methodologies along with their tools that help, developing OCR systems are designed for the purpose of data collection, segmentation of characters and feature selection. Using the SVM classifier, 94% accuracy is achieved for Telugu scripts.
The developed tool is used to accomplish the task of certain phases of OCR system and it works only for Telugu language.

Indira et.al. [28] discussed the classification and recognition of Hindi Vowels and Consonant characters by dividing the task of OCR system into five stages. Three subgroups are made from all the characters such that one group contains the characters without vertical bar, one group contains the characters having the vertical bar at right side and the other group contains the characters having vertical bar at middle of the character. Back propagation Neural network classifier is used for each group separately. The recognition accuracy achieved is to be of 95%, when the training data used for testing and 76% is achieved for untrained data. Even though it considers only basic symbols (i.e. not considering conjunct characters) closely resembling characters that are misclassified by the neural network because of considered structural features.

Yadav et.al. [29] reported that resolution of the document image should be of 100dpi to 600dpi. Three new feature extraction techniques are proposed: histogram of projection based mean distance, histogram of projection based on pixel value and vertical zero crossing. For the purpose of classification and recognition, the conventional back propagation neural network with two hidden layers architecture is designed. Comparative analysis among the different combinations of feature extraction and classification techniques is represented for Hindi vowels and consonants. Promising result is achieved in the proposed technique and character level rate of recognition is 98.5% is justified by them that it identifies 383 out of 385 characters.

Chaudhuri et.al. [30] proposed a complete OCR system for Bangla language. As a part of the pre-processing of the document image, digitization is achieved by histogram based thresholding approach, skew detection as well as correction is done by connected component labelling. Segmentation of line, word and character is performed by analysis of horizontal and vertical pixel analysis. The feature of the characters are extracted by finding eight strokes and considering the dots filled by a circle. They have used a tree based classifier for recognition of basic characters as well as modifiers. However, the compound characters are classified into two levels. At first the feature based tree classifier is used to separate them into small groups and then a template matching classifier is used to recognize the characters in each group. It concludes with the result obtained by testing the Bangla
characters of different font size and reported that the finest rate of recognition is 96.70% for the font size 14pt.

Mahmud et. al. [31] proposed an OCR system for isolated and continuous multi-font Bangla characters. Like other researchers, the pre-processing is performed in accordance with a normalization of individual segmented character which is done to make all of uniform size. The feature vector is obtained for all images by computing the freeman chain code. The classification is performed using a feed forward neural network of 4 layers (1 input, 2hidden, 1 output). It is reported to achieve a 98% accuracy for isolated characters.

Hasnat et.al. [32] improve the Bangla script processing power by using free open source OCR engine Tesseract. It is considered as one of the most accurate and powerful engines that works for different languages, which is maintained by Google. They have considered 340 characters of the script that includes vowels, consonants, numerals and compound characters. It is observed to be of 93% accuracy for printed document by training the engine.

Bag et. al. [33] present a review of intensive research on character recognition system for two most popular scripts like Bangla and Devanagari. It contains the summarization of most of the published papers and gives a detail picture of the methodologies adopted as well as the recognition rate reached by the systems.

Bag et. al. [34] address the recognition of compound characters of the Bangla script only. Shape of the compound characters are complex in nature as more than one characters are combined to form a single character. Authors propose a feature extraction technique based on the convex shape of the skeleton image characters. This method is named as Topological feature. A matching score is computed (similarity measure) with the previously stored template images. It could be able to achieve 90.20% of recognition rate for printed character using the proposed approach.

Kompalli et. al. [35] have proposed a graph based method for segmentation of characters from the image document. A graph is obtained from topological structure of a character image. Here Neural network classifier is used for classification and finite state automation is used for the construction of words.
Dongre et. al [36], Jayadevan [37] investigated the research direction of the Devnagari script. They broadly described the techniques used for each phases of the OCR system. The feature extraction methods are classified into three major groups: Global Transformation, Statistical and Geometrical and Topological features. Classifiers adopted by various researchers are categorised into: Template matching, Statistical Techniques, Neural network and Support Vector machine. They also analyse the limitations of methodologies which can be classified based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or hand-written).

Also extensive research of OCR is extended to Tamil [38], Telugu [39], Malayalam [40], Urdu [41], Kannada [42] and Gurmukhi [43] scripts.

2.5.2. Odia Language

Odisha is a state having rich heritage of novels and manuscripts. Many of these are written hundreds of years ago. Hence, these novels and manuscripts need preservation with easy accessibility to serve the society. This can only be possible by implementing an OCR system that efficiently and effectively converts Odia language documents into electronic form.

Many researches have been conducted to design an efficient Odia OCR, design of an efficient and robust Odia OCR system. It is very challenging as the Odia script includes a large number of base-characters in its alphabet set and again the compound characters are formed by joining two or more symbols that further complicate the problem. So far, the research works proposed for Odia language mainly deal with recognition of basic Odia characters. Research work published on Odia character recognition is very limited as compared to different other Indic languages.

Chaudhuri, et. al. [44] uses the basic character of Odia language and applied some conventional techniques for character segmentation as well as for pre-processing task such as skew correction, zone detection, normalization etc. There are many techniques designed for skew detection and correction such as horizontal and vertical projection profile, nearest neighbour clustering, Hough transform, Fourier transform etc., but for Odia script, Hough transform technique is recommended. He suggested a feature extraction technique borrowed from the concept of water overflow from reservoir, and run-length feature. For
classification, it used tree classifier and could achieve 96.3% for character level recognition. The boundary tracing method used for recognition, fails to identify the similar shape characters of the Odia script. Also the proposed segmentation approach is unable to segment touched characters, found in the images taken from very low-grade quality old printed documents. In general, Odia documents may also contain the fonts of different size, which is not been considered in the proposed system by the author.

Mishra, et. al. [45] also tried for basic characters recognition. Each segmented image is resized to 16×16. Ten structural features are extracted by measuring their Dimensions, Storage details, Histogram, Pixel counts etc. (such as Upper Part Circular, a vertical line on the right most part, existence of Holes, Number of holes). Recognition is achieved with respect to the matching score of the Feed-forward neural network and the percentage of recognition has not being reported. Finding of the above measured characteristics from scanned digitized image of Odia printed document is very difficult as it is influenced by the noise. During scanning, noise may occur because of existence of dust particle over the surface of scanner, low hardware quality of scanner as well as printer, low quality of paper and image of ancient document etc.

Tripathy [46] considered only basic characters. Zernike moments as well as Hu moments are extracted of the characters and also reported the difficulties of Hu moments. Instead of using any classification technique, this work reports the reconstruction of the characters with the order of 30, and made a comparison with both of the techniques. But the result is limited to few basic characters. For reconstruction of the characters according to this approach, there is a need of library of 1000 number of characters of each glyph in different size and different type. However, it is found to be very much difficult to get a library, as limited fonts are designed and available of this script.

Mohanty et. al, [47] used SVM and K-NN techniques to study the recognition of only basic symbol set of Odia script. They considered the longest run feature of the characters. To improve the classification accuracy and to discard unrequired features (feature values less than a threshold), Mohanty et. al. proposed a wrapped based feature weighting algorithms, which finds globally optimal feature set. In their proposed evaluation scheme, concluded with accuracy rate of SVM is 98.9% as compared to 96.47% with K-NN. However, they reported the choice of kernel for SVM as well as time and space
needed by it as biggest limitation. As some Odia scanned documents contain bold and italic font style, a robust OCR should consider those with the data set which has not been considered by the author.

Mohanty [48] has presented a methodology of an unsupervised, self-organizing, competitive learning feature map named as Kohenn neural network (KNN). The network recognizes the characters through its clustering and learning capacity. Each character pixel patterns (Gray scale values) are mapped to 8 × 8 matrix. The KNN uses two layers each with 8 × 8 linear units. However, the result reported in this article considers only 5 characters of the script.

As few documents are found in English and Odia script, Mohanty et. al. [49, 50] made an attempt to develop a Bilingual-OCR. At first, the segmentation is performed at script level. Horizontal and Vertical projection profile of the characters are used as features and classification is performed by SVM separately for each script. As some symbols of both the scripts look alike, the rate of recognition decreases. Also a post processing needs to be designed as recognized characters of both scripts are included in a single editor.

Mitra et. al. [51] have studied the directional features extracted in two different approaches. The simple images of size 48×48 are subdivided into equal sized 8×8 zones and from each zone 4 directional features i.e. a total of 144 features, are extracted. To overcome the difficulty for complex images, the zones are chosen based on the directional stretch of character pattern. For the purpose of classification, it uses the C++ library LibSVM. The training data set consists of 2000 images and the testing is performed over 936 images. It shows that 95% of recognition accuracy can be achieved. The considered directional features are not able to distinguish composite characters of Odia script as they provide almost identical values and the author also reported that it cannot be claimed to be a full-fledged OCR system for printed characters.

Nigam et. al. [52] proposed a feature extraction technique based on curvelet transformation which is based on parabolic scaling law. This proposed approach uses different font size images. After 2-level of thinning and 2-level of thickening of morphological operation performed over the images, SVM classifier is being used. The experimental result reported in this article achieves 94.7% of recognition. The
computational cost of a curvelet transform is approximately 10–20 times that of an Fast Fourier Transform (FFT) and for an image of $n \times n$ size complexity is of $O(n^2 \log n)$.

Kumar et.al [53] divide each character into $n$ Zones and uses the zone centroid method to extract global features such as number of horizontal strokes, vertical strokes, angular strokes, aspect ratio, open loop, closed loop etc. They have also divided each character into $n$ segments and used fuzzy logic based feature extraction technique to find local feature such as number of cross point, centre of gravity of end points etc. The classifier SVM is practised along with the rule based Ant miner algorithm for classification and recognition. This paper uses 12 characters of training pattern and accomplishes the recognition rate up to this mark 92.42%-97.87% by performing the test, considering same set of training samples. As composite characters of the script are complex in shape, important information of which can be lost by dividing it into zones and therefore the author only considers the basic characters.

Senapati, et. al. [54] carry out the analysis for one important phase of OCR system named as segmentation for the printed Odia document. At first, each line of the document is separated into three zones named as upper, middle and lower zones. Afterwards, segmentation is achieved in a hierarchical manner that is first at line level, then word and then at character level from each zone of the printed document. That paper reports 99.3% accuracy which can be achieved for line and word levels. However, character level segmentation can only be able to obtain at 86% accuracy, and also the complicacy of segmentation for Odia matra is reported.

Pati et. al. [55] also do the analysis of Vowels and Consonants of the script. A simple histogram based grey scale binarization technique is adopted. The line, word and ultimately character level segmentation of the image document is done by dividing each into three zones. Aspect ratio of the images is extracted to normalize the images. The moment based features are extracted here and use the k-NN classifier for recognition is used. It reports that 72.27% of accuracy being achieved. The moment based features are the statistical measures of the images, moreover the authors have not used any statistical measuring techniques to prove any significant differences of extracted features among different characters.
Mohanty et. al. [56] describe the character recognition process for the printed document of multi-font, i.e. with the consideration of Roman and Odia scripts. The segmentation of character is done by performing horizontal and vertical projection. Aspect ratio of the characters is taken into consideration to extract the feature as well as distinguish among the two different scripts. This paper did not mention about the data set used for training and testing and also about the recognition accuracy.

Mohanty et. al. [57] also extended the work of Bilingual (English-Odia) OCR with the help of different classifiers. They have accomplished the segmentation by dividing each line into three zones and considered aspect ratio as the distinguishable feature. This article uses different classifiers as (k-NN, CNN, SVM) and reported SVM to be the best which gives 98% of accuracy. However, this work is reported for isolated, basic (i.e. Vowels, Consonants) alphabet of the script. Furthermore, the testing result is mentioned for both the script separately. Some efforts also are made towards the development for Odia handwritten numerals and characters by researcher [58, 59].

2.6. Summary

Even though, it was a herculean task for researchers on OCR to develop a full-proof OCR application for hundreds of languages with thousands of characters, the sheer advantage of OCR to enable them to preserve lakhs of manuscripts and documents of importance for posterity with very little space and with option to recover the saved documents in no time, has attracted the researchers across the globe for a full-proof OCR application. OCR application, which converts characters in any language to electronic signal, helps transferring such data across the world in no time and ensures seamless assimilation of knowledge.

OCR application undergoes through many technical stages to convert the characters to electronic signals which are technically described as Digital Image Processing, Computer Vision, Pattern Recognition, Natural Language Processing, and Artificial Neural Network. Similarly, the World today has seen many reputed companies promoting applications like ABBYY FineReader version 12, OmniPage Professional version 18, Readiris Pro 12 and NovoVerus etc, which are very efficient in their field.
But, in spite of early attempt from nineteenth century and onset of electronic revolution in twentieth century, the world is yet to see a fool proof OCR application with hundred percent accuracy. Still, attempts are being made by the Researchers to develop more efficient OCR applications even though major languages had seen near-perfect OCR applications. But other peripheral languages like the present language i.e. Odia, is yet to see an OCR application of significance. Even though, many researchers have attempted and are still attempting to put in place a perfect OCR application for Odia language, they are far away from their goal.

The Odia language, spoken by 40 million native people and stands at 37th rank in terms of native people speaking the language in the World, and has ancient past, which is found in Dhauli edict in 3rd century B.C. It has a chequered history having a vast array of contribution to literary world. Since the time of Dhauli edict, Madala Panji, Sarala Das’s Vilanka Ramayan, the literary world of Odia language has been richly glorified. But, it is very ironical that in spite of wealthy of literary resource the language is yet to see an OCR application to preserve all these monumental works for posterity and propagate the same world over.

The Researcher in the present chapter has extensively surveyed the literature and urged upon the urgency of adopting an OCR application to Odia language.