Abstract: Telugu Handwritten Character Recognition System (THCRS) is a field of Optical Character Recognition (OCR) and Pattern matching. Variations in handwriting from person to person makes the translation of the handwritten text into digitized text format and recognition of the same by a machine is still a tough task and challenging in today’s much-advanced world. If user’s Telugu handwriting style is not clear then the situation becomes worse when two characters are intermingled or overlapped and appears as a single character. So, converting Telugu handwritten characters in digital text format are quite difficult when compared to other languages, since Telugu characters have many curves and strokes in nature. The Telugu handwritten character extraction process is hierarchically divided into an acquisition, pre-processing, segmentation, feature extraction, classification, recognition and recognition and post-processing. The main goal of extraction is to take out the common features of patterns of handwritten characters, recognize those characters and produce the same in digital format. We have done some survey on previously developed HCR systems and explored in this paper. This paper presents a brief idea on different methods which are capable of recognizing different handwritten texts or characters taken from different individuals with great accuracy.

Keywords: Telugu characters, OCR, THCRS, Pre-Processing, Segmentation, Extraction, and Recognition.

1. Introduction

OCR is a very much interactive tool to improve the information exchange process between user and machine [1]. It can be used to convert machine printed or handwritten documents into editable text format. Major levels in an OCR System are given in Fig: 1.1
the primary language in Andhra Pradesh, Telangana and Pondicherry (Yanam) state in India. Telugu script has high complexity among all other Indian scripts, as it has largest number of vowels and consonants, Complex Composition Rule. Telugu Script is a combination of 14 Vowels, 36 Consonants, and 3 Special Characters which are shown in Figure 1.2, 1.3 and 1.4 [6] [7].

Fig: 1.2: List of all vowels in primary and secondary form.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Primary</th>
<th>Secondary</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>ఆ</td>
<td>ఆ్</td>
</tr>
<tr>
<td>i</td>
<td>ఇ</td>
<td>ఇ్</td>
</tr>
<tr>
<td>u</td>
<td>ఉ</td>
<td>ఉ్</td>
</tr>
<tr>
<td>a</td>
<td>ఐ</td>
<td>ఐ్</td>
</tr>
<tr>
<td>i</td>
<td>ఒ</td>
<td>ఒ్</td>
</tr>
<tr>
<td>u</td>
<td>ఔ</td>
<td>ఔ్</td>
</tr>
<tr>
<td>a</td>
<td>ఎ</td>
<td>ఎ్</td>
</tr>
<tr>
<td>i</td>
<td>ఏ</td>
<td>ఏ్</td>
</tr>
<tr>
<td>u</td>
<td>ఐ</td>
<td>ఐ్</td>
</tr>
<tr>
<td>a</td>
<td>ఐ</td>
<td>ఐ్</td>
</tr>
<tr>
<td>i</td>
<td>ఐ</td>
<td>ఐ్</td>
</tr>
<tr>
<td>u</td>
<td>ఐ</td>
<td>ఐ్</td>
</tr>
</tbody>
</table>

Fig: 1.3: List of all consonants in Telugu

<table>
<thead>
<tr>
<th>Consonant</th>
<th>क</th>
<th>च</th>
<th>छ</th>
<th>ज</th>
</tr>
</thead>
<tbody>
<tr>
<td>क</td>
<td>क</td>
<td>क</td>
<td>क</td>
<td>क</td>
</tr>
<tr>
<td>च</td>
<td>च</td>
<td>च</td>
<td>च</td>
<td>च</td>
</tr>
<tr>
<td>छ</td>
<td>छ</td>
<td>छ</td>
<td>छ</td>
<td>छ</td>
</tr>
<tr>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
</tr>
</tbody>
</table>

Fig: 1.4: List of all consonants-vowels combinations.
There are huge amounts of old handwritten documents in Telugu script available. Researchers proactively have given attention towards the amazing language for OCR. On this regard, various methodologies have been used by researchers based on Hidden Markov Models (HMM), Elastic Matching, Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Raspberry Pi [8]. Now a day’s gadgets like small hand-held mobiles, tablets, smart watches are rapidly growing in the market. In all these types of gadgets, User Interface Screens (UIS) are designed in their local language. As of now there are many tools available to design UIS in English language but very few are available for Telugu language [9]. Many research investigations are carried out on image-based hand gesture recognition that needs the text input system [10]. Some studies specify on non-touch input methods of computer systems and some other are focused on support for the deaf and dumb people. These methods do not need keyboards, mouse devices, and body-worn devices, e.g., cyber-gloves but they need image scanning and capturing devices like cameras [11]. These methods can be commonly applied as most of the mobile gadgets have an inbuilt camera. Although, voice recognition also supports non-touch input, it has some drawbacks like being vulnerable to surroundings noise, privacy problems, and problems of mispronunciation and a voice disorder of users [12]. According to the World Health Organization (WHO), approximately three hundred and sixty million people worldwide having hearing loss and seventy million are deaf. So, some studies were also done on sign language and eye gestures [13] [14].

Digital text data can be extracted from optical scanning by OCR or intelligent word recognition (IWR) from the scanned copy of handwritten text document. Optical character recognition is the translation of handwritten or typed text into an electronic format, which can be stored, interpreted and processed by a machine. This hand written character recognition process is emerging field and part of pattern recognition too. As mentioned in Fig: 1.1 many functions are to be carried out to get right output from the text document. Each process has its own pros and cons where cons need to be fixed to step ahead for extraction process [15].

**Categorization of Character Recognition Systems**

Character recognition (CR) has been growing rapidly due to the huge requirement of online and offline documents scanning and parsing, since late sixties, Categorization of CR is given in Fig: 1.5. There are number of research studies available which contain either complete or partial information regarding the character recognition. The curvy, single, multi-stroke and complex nature of characters for recognition from printed/scanned document source creates more challenges in this field of research.

Fig: 1.5 Categorization of Character Recognition Systems.
Character Recognition Technologies

1. Optical word Recognition (OWR): Here output is typewritten text i.e. one word at a time.

2. Intelligent Character Recognition (ICR): It outputs the handwritten print characters or cursive/curvy text i.e. one character at a time, this generally involves machine learning (ML) techniques.

3. Intelligent Word Recognition (IWR): It outputs handwritten print characters or curvy text one word at a time, generally used for the languages where characters are not separated in cursive script.

All the above-mentioned types are represented in hierarchical pictorial form in Fig: 1.6. The flow control statement can be given as,
If ((Better input image quality) && (Better Character Recognition Methodology))
then
Better character recognition accuracy.
else
Low character recognition accuracy.

Fig 1.6: Different types of OCR

Attributes of Indian scripts

Indian languages have mainly common alphabet, though they use many forms to express it. The alphabet has vowels, consonants and a rich combination of them gives a wide range of character forms. Indian scripts are not case-sensitive. Many Indian scripts are written from left to right associatively except Urdu script which is written from right to left associatively [16]. In general, some scripts like Bengali, Hindi, and Assamese have horizontal and vertical linear features, while others like Telugu, Tamil and Malayalam have complicated curves and strokes. In India, there are eighteen official languages, namely Assamese, Bangla, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Rajasthani, Sanskrit, English, Gujarati, Hindi, Konkani, Nepali, Tamil, Telugu and Urdu [17]. Now a couple of other languages are also added to the list.
One hundred rupees
एक सौ रुपये

Fig: 1.7 Examples of 12 Indian scripts: 'Top to bottom: English, Devnagari, Bangla, Gujarati, Kannada, Kashmiri, Malayalam, Oriya,Gurumukhi(Punjabi),Tamil,Telugu, Urdu. All lines have same meaning.

Many researchers have done some studies on this emerging field of handwritten character recognition system and we have arranged all those studies below in ascending year wise.

**Some previous research studies on the HCR System from the Period of 1991 to recent years**

*Turk M and Pentland A,* [18] had developed a feature extraction concept that utilized Eigenfaces for Recognition. *Jacobs R A et al.,[*19*] proposed a model for modular neural architecture which has multiple expert networks and a porting network. In class modular model one high complexity task spreads into a number of multiple subtasks and each subtask is assigned to an expert network. *O. Matan et al.*, [20] studied and modelled segmentation concept on the basis of vertical cut which analyzed vertical projection of the picture in order to segment the ZIP codes of places or locations. *Cheriet M and Suen C Y,* [21] focused on pre-processing concepts as it is necessary to work out document image analysis operations prior in recognizing the text in scanned documents. *Cardot H et al.*, [22] proposed a static signature verification system by co-operative neural network architecture (CNNA). During scanning process of the text some noise like gaps and bumps in lines, disjoint lines were found. *Serra* [23] carried out research on morphological filtering where some noise removal methods were discussed. *Bimbo a D et al.*, [24] studied OCR which recognizes low-quality images by deformation of elastic templates. *Chen M Y et al.*, [25] developed a method which recognizes offline handwritten words using a Hidden Markov Model (HMM) type stochastic network. *Trier D and Jain A K*, [26] focused on goal-directed evaluation of binarization methodology which is a spatial domain approach required for compression. This approach applies a concept called thresholding which minimizes the storage requirements and helps in increasing the speed of processing by converting the color (R-RED, G-GREEN, B-BLUE) or gray-scale pictures to binary by using a threshold value. *Bramall P E and Higgins C A* [27] derived a method that recognizes the cursive type of handwriting from scanned documents based on human reading Models. *Bauer F and Wirtz B* [28] designed a model based on parameter reduction and personalized parameter selection for automating signature verification process. *Chen H et al.*, [29] proposed a piece-wise linear modulation method for handwriting character recognition. *Dehghan M* [30] defined a model that performs the task like signature verification using shape descriptors and multiple neural networks. *Bouletreau V et al.*, [31] shaped a model to recognize handwriting text characters and signatures of individuals to analyze the unique nature of writing, having application in the digital forensic analysis. *Kim et al.*, [32] found out a model for handwritten character recognition systems where the post-processing stage of the handwriting character recognition system was discussed. *Coates M et al.*, [33] developed an automatic reading of cursive scripts using a reading method and perceptual method. *Morita M et al.*, [34] defined a method which used mathematical morphology and weighted least squares to do corrections of handwriting baseline skew. *Lee S W and Kim S Y* [35] brought out a new type of cascade neural network (NN) to train the spatial dependency in connected handwritten numeral characters, which actually extended from the multilayer feed-forward neural network to improve the discrimination and generalization power. *Nishimura N et al.*, [36] designed a model which is for offline character recognition using
HMM by multi-directional feature extraction and voting with bagging algorithm. Atul Negi et al.,[37] proposed an OCR system for Telugu script. The two important features of this system are practicability and simplicity. The connected component algorithm used for segmenting words into recognizable units and the template matching used to recognize the components. Marti U V and Bunke H [38] studied the IAM-onDB an On-line English sentence database acquired from handwritten text on a whiteboard for off-line handwriting character recognition. Tomai C et al.,[39] focused on transcript mapping for recognizing historic handwritten characters of document images. Bunke H [40] highlighted the state of the art in offline Roman cursive handwriting recognition where the input was an image of a numeral or a digit, a word or some text or pattern and the system produced an ASCII output neuron which shows the cluster,

other localization methods for various Applications, Considered 7 applications for study. N Prameela et al.,[61] proposed a very comprehensive and practical OCR system for Telugu language. This system is shape and font dependent that requires pre-processing and feature extraction.
Conclusion
It has been observed that researchers carried out different types of approaches during the last twenty plus years and suggested many ways for pre-processing, segmentation, recognition and post-processing. The studies during the recent years are based on diagonal based feature extraction, zone-based method, distribution based method, sliding window, freeman-chain-codes, and Eigenvalue. Many studies were conducted on different scripts like English, Devanagari, Bangla, Gurumuki, Tamil, Telugu, Chinese, Japanese, Kanji, and Arabic. This paper is to generalize the survey and to brief the achievements of the previously focused work by researchers. Many studies were on limited data set or data set for some characters in Telugu language. Some have high time complexity for recognition. Finally, research on Telugu handwritten character script recognition can be moved ahead by taking high data set for a maximum number of characters and by using hybrid character recognition algorithms with best case time complexity.

References
