Online Handwritten Indian Character Recognition and its Extension to Telugu Character Recognition

Srilakshmi Inuganti, Rajeshwara Rao Ramisetty

Abstract: ***In this paper, we portray the cutting edge of Online Handwritten Character Recognition (OHCR) as for preprocessing strategies, include extraction and acknowledgment techniques, alongside different information accumulation gadgets accessible. OHCR is the method of recognizing characters by a machine while the user writes, in which the handheld devices record (x, y) coordinates of the track of the character. With the advent of handheld devices, there is a great attention towards OHCR of regional languages. The datasets available in various Indian Script are also illustrated, as benchmarking database is very crucial for any research. Finally, future scope in Telugu OHCR is mentioned, as very less progress in Telugu script when compared to remaining Indian scripts.***

Index terms: OHCR, Online Handwritten Character Recognition, OHCR a Survey, Telugu.

I. INTRODUCTION

With handheld devices reaching new heights of popularity every day and becoming almost With handheld devices reaching new heights of popularity every day and becoming almost indispensable in our busy lives, digital pens become a great alternative to keyboards, especially in case of PDAs, Hand Held PCs and high end mobile devices. A digital pen captures the handwriting of a user, converts handwritten information into digital data, enabling the data to be utilized in various applications. In this context Handwritten Character Recognition (HCR) is an immediate challenge in the area of pattern recognition. HCR can be classified into Online Handwritten Character Recognition and Offline Handwritten Character Recognition. Handwritten Character Recognition is the task of identifying character written by a machine while the user writes, in which transducer required for capturing dynamic handwriting information. The dynamic information contains numbers, order, length, writing direction and speed of stroke and some devices record pressure information also (i.e. At pen tip). A stroke is the writing form pen-down to pen-up. Offline HCR is a sub category of Optical Character Recognition.

In offline HCR character is recognized after completion of writing. Offline HCR takes a raster image from digital input source and converts into binary image, so that image pixel values are either 0 or 1. The progressive study of handwritten character Recognition shows online HCR has many advantages over offline HCR. Offline data is not associated with temporal information. It only represents the final result as an image. So knowledge about the character is less. Online data are associated with temporal information, so that accuracy is high in adverse to offline. Online data are highly interactive. Hence, errors can be debugged immediately with repeated tests. Memory required for online representation is very less. Even though many years of research in handwriting recognition [1,2], very less has been made towards Indian languages. OHCR is more realistic for Indian languages which have huge character set.

II. CHARACTERISTICS OF INDIAN SCRIPT

India is a limitlessly multilingual nation with 22 official dialects perceived by constitution. Likewise, several non-official dialects are being used, every one with an assortment of tongues. Directly there exist 11 noteworthy contents, which are as of now utilized in India: Assamese, Bengali, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Odiya, Tamil, Telugu, and Urdu. Of these, Urdu is acquired from the Persian content and is composed from appropriate to left, staying ten other, composed from left to right, protected from the early Brahmi content and are likewise alluded to as Indic contents. Albeit Indian dialects utilize diverse structures to express it, they have basically a typical letter set. The idea of upper or lower case characters is absent in Indian contents.

The advantages of high advancement in data innovation (IT) area presently can't seem to permeate down to the grassroots dimension; There is a mind blowing development in Internet use in India (28 million clients), it

Revised Manuscript Received on December 22, 2018.
Srilakshmi Inuganti, Lecturer in Computer Science Government Degree College for Women, Srikakulam
Dr. Rajeshwara Rao Ramisetty, Professor in Computer Science and Engineering, JNTUK, UCE

represents 2.72% of the Indian populace, represents 2% of the world Internet clients [3]. Notwithstanding monetary factor, this difference has come about on account of dreary word handling of Indian Languages with QWERTY console utilized for English letter sets. Indian dialects have a mind boggling character unit or Akshara that is the nuclear etymological unit. An Akshara can be framed with 0, 1, 2, or 3 consonants and a vowel. Thus Akshara is essential unit of word. The run of the mill types of Akshara are V, CV, CCV and CCCV, therefore have a summed up type of C*V. Subsequently the fundamental units of character of content O (102), these units framing O (104) number of composite characters. An agreeable arrangement is utilize phonetic interfaces, for example, discourse and penmanship recognizers to PC and different gadgets, which assume an essential job in giving financial development in rustic zones.

III. FRAMEWORK OF OHCR

The block diagram OHCR illustrated in Figure1.The details of each step are described in the following paragraphs.

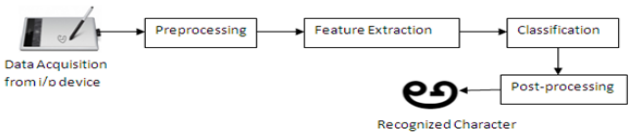

Fig. 1: Steps in OHCR

a. Data Collection

By utilizing web based penmanship acknowledgment programming like digitizers transformation of the content at the same time with the client's composition is conceivable. These digitizers like PDAs make use sensors to record development of the info gadget like a stylus pen. In this when the pen tip is in contact with the screen, the sensors are activated. At the point when the contact is broken, the sensors are naturally killed. The procurement interface records a grouping of (x, y) - facilitates speaking to the area of pen tip and twofold esteem shows pen up/pen down, the directions are recorded just period when the pen is in contact with the interface. This period is known as stroke

Commercial Products Available

Various penmanship acknowledgment programming sellers consider the sort of computerized gadget that can have their projects. This is on the grounds that the utilization of penmanship acknowledgment innovation is quick turning into a standard for cell phones, however for other gear also. PDAs, workstations, tablets, and cell phones are nevertheless a portion of the advanced gadgets which utilize penmanship acknowledgment programming in their activity. Now a days Digimemos are also used which operate with only normal pen. In following paragraphs some commercial products available are described.

The Wacom Intuos Pro is a tablet with stylus [4]. It is available in multiple sizes and it's pen technology is pressure sensitive and cordless. This product acts as an input device by connecting it to the computer via USB. The pressure level is of 1024 (pen tip only).It has a resolution of 100 lines/mm (2540 lpi).The reading speed of pen is 133 pps. This sleek tablet is compatible with a high range of operating systems such as the Windows 7, Mac OS, Windows XP, Vista SP2, and so on.

A tablet PC is an exceptional notepad PC with a digitizer tablet and stylus.It allows user to write on screen, the operating system recognizes the character. iBall Pen Tablet offers users the flexibility to write, interact with the graphical user interface and performs other input related tasks in a very user friendly manner [5]. It offers 1024 pressure levels.

The ACECAD Dig memo [6] is a standalone device that captures and stores everything you write without the use of computer and special paper. When connected to a PC, it offers on-line handwriting functions which can synchronize your writing on paper with its digital page in its software window. It is very portable, so that people can easily operate while standing or sitting. People feel as comfortable as with a regular pen on paper. So that data can be easily collected from computer novice also.

HP labs Data Collection Tool (DCT) provides user interface by which handwriting samples from different writers can be collected [7]. Given the script and the data elements to be collected, the writer can give their handwriting samples. The current version of DCT is supported on a Tablet PC with Microsoft Windows XP Tablet PC edition and Desktop PC with Microsoft Windows XP operating system. Figure 2 shows some of the commercial products available.

Fig. 2: Some Commercial Products

Standards for Online Data Representation

The following are the main standards for representing online-data

UNIPEN Format

The UNIPEN format is a common data format, easy to exchange [8]. Users can easily convert their data to UNIPEN format, collect new data directly. The data were developed and tested in collaboration with many industrial experts. The strokes are recorded from .PEN_DOWN to .PEN_UP.

Digital Ink Markup Language

InkML is an markup language to represent "ink" data with an electronic pen or stylus [9] based on XML. The

suggested determination was distributed by the World Wide Web Consortium (W3C) in September 2011.By using single <ink>element entire content of an InkML document is enclosed. The<trace> is the basic data element in an InkML file, where a trace is used to represents a sequence of contiguous ink points, in which each point captures the values of particular quantities such as the X and Y coordinates of the pen's position.

UPX

UPX, an XML-based successor of UNIPEN. It addresses the limitations of UNIPEN and InkML [10]. It allows the user to easily add specific information to ink files to suit the needs of the application.

3.1.3 Dataset Available

IWFHR 2006 Online Tamil Handwritten Character Recognition Competition data set can be used. It contains isolated character samples of 156 Tamil characters written by native Tamil writers, including school children, university graduates, and adults from the cities of Bangalore and Salem. It is collected by using an HP tablet and in UNIPEN format [11]. It consists of 50,385 training samples and 26,926 test samples across 156.The dataset hpl-telugu-iso-char-online-1.0 contains samples of the 166 character classes collected from different writers on ACECAD Digimemo (A4 sized) using an AcecadDigi memo DCT application. It consists of 50,385 training samples and 26,926 test samples for 166 characters. HP dataset also available for Devanagari characters. Computer Vision and Pattern Recognition Unit of the Indian Statistical Institute developed ISI handwritten character databases , which consist of several items which includes (i) numerals of Bangla, Devanagari and Odiya scripts and (ii) basic, vowel modifier and compound characters of Bangla script [12]. The databases can be utilized for research purposes. These samples are collected by using WACOM Intuos 2 tablet.

b. Preprocessing

Before applying input to the system to get correct recognition result, data need to be pre-processed. As the data collected in real environment, it can be noisy and inconsistent.

Normalization

Usually the recognition rate is high, if we normalize the character with respect to the width and height, along with a starting point.

Smoothening

Smoothing is performed to reduce the jitters in input obtained from the hardware or hand motion.

Removal of Repetition of Points

Sometimes input data contains duplicate points and does not contain any useful information for classification.

Interpolation

Interpolation is the prerequisite for applying Re-sampling. Interpolation generates missing points, usually with the constraint that distance cannot be more than a certain threshold.

Re-sampling

Re-Sampling is performed to normalize input character to a constant number of points which are at equal distances. When the character has multiple, each stroke is resampled to preserve its ratio with the character.

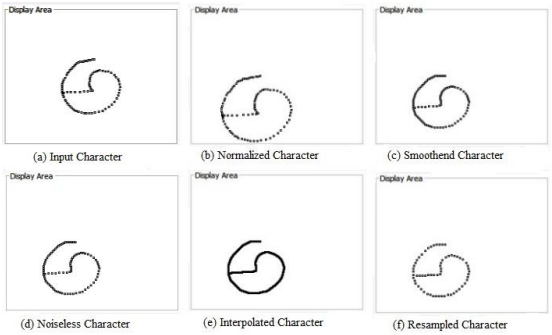

Fig. 3: Pre-Processed Character

Even though preprocessing enhances recognition accuracy, excessive preprocessing is undesirable because it may result in loss of valuable information. The preprocessed character shown in the Figure 3.

c. Feature Extraction

Feature extraction starts with measured data and builds features, which are informative and non-redundant. These features extracted should maximize inter-class similarity and minimize intra-class similarity.

Fixed Length vs. Variable Length Feature Vector Representation

Features of online handwriting units can be fixed length or variable length [42]. In fixed length representation, predefined length of feature vector is extracted from stroke. Fixed length representation uses reserved space large enough to accommodate larger data. These features include direction information, direction feature, structural features, etc., Variable length feature vector is suitable for complex stroke and structure variations. This representation is useful for template matching like Dynamic Time Wrapping (DTW) and Hidden Markov Model (HMM) based approaches. Fixed length representations are useful in Support Vector Machine (SVM) and Principal Component Analysis (PCA) based approaches [58,64].

Local vs. Global Features

Contingent upon the granularity at which highlights are extricated, they are named neighborhood and worldwide. Neighborhood highlights are removed at a point of the stroke. The most widely recognized nearby highlights are x-y arranges, neighborhood course includes, for example The overall vector of two nearby focuses [61]. Worldwide component is characterized as a relative vector between any subjective focuses [13]. These highlights are extricated at a stroke level or sub stroke level. Instances of worldwide highlights are minutes, Fourier descriptors, projections, and so on. Worldwide highlights are great at catching by and large data. They don't function admirably with comparable classes that have minor varieties. Nearby highlights are extricated at each point, useful for between class detachment. The highlights proposed in the writing so far are either neighborhood or worldwide which neglect to catch basic data about the character. A mix of nearby and worldwide highlights has been proposed to catch

neighborhood and worldwide varieties [55].

d. Classification

A number of different models have been applied to Indian OHCR. Different models of online handwriting recognition are illustrated by Bellegarda in [14]. These recognition models are Motor Models, Structured Based Models, Statistical Models, and Neural Network Models. Work in each of the above mentioned methods is illustrated in the following sections. The merits and demerits of each of these models given in Table 1.

Motor Methods

Motor models [15-17] are a technique commonly used in what is known as Analysis by Synthesis in which models of stroke segments are created along with rules for connecting them to form characters. Motor models represent these stroke segments as parameterized models of the motion of the pen tip, simulating the physical properties of human hand motion.

Structured Based Models

In Structure Based Methods different examples of strokes are considered as primitives [18]. The distance of test pattern with reference pattern is calculated. The distance measures can vary from the Euclidian distance to Mahalanobis distance. Structure based methods are weak at collecting data, but good at recognizing variations. The elastic matching approach followed in [19,20], which works on sequence of sample points directly by comparing the alignment of input pattern with reference pattern. A direction string approach is described in [21], in which each stroke is represented in the form of direction followed. DTW method is followed in [22]. DTW compares online trajectories of the coordinates; trajectories include temporal and spatial information. If the character is represented as graph, graph matching algorithms can be applied to classify the character. Delaunay triangulation features are used in [23].

Statistical Models

These methods are probabilistic and need powerful calculators and considerable calculation time. This character is classified by selecting the class which is most probable or has a minimum amount of classification error. Well known probabilistic model Bayesian decision rule is applied for OHCR in [24]. A probabilistic discriminate model Conditional Random Fields is applied in [25]. Another popular statistical method is HMM, the success of which motivated towards the application of HMM to OHCR [67]. HMM is trained from each stroke class using the observation sequence obtained when stoke is written. HMM allows variable length feature vector. Supervised learning method SVM is also another widely used OHCR method [42]. Decision trees can also be applied for classification, where prior probabilities can be used [26].

Neural Network Models

Neural Network methods for OHCR are gaining popularity, because of their performance in other areas. Multi Convolution Neural Networks are applied in [27]. Combination of HMM and Time delay Neural Networks have shown good performance for cursive script recognition in [28]. A Recurrent Neural Network's approach is applied in [29]. It has the ability to make use of previous context.

Table 1: Review of different techniques of online handwriting

Techniques	Merits	Demerits
Motor models	Uses advantage of pen dynamics	May lack robustness when writing style variations are large
Structure Based models	Works well for writer-dependent data	Does not work very well for writer-independent data, Recognition time rises linearly with the number of training examples
Statistical models	Temporal relations are very well modelled	Requires a large amount of training data
Neural Network models	Less Classification time	Temporal relations are not very well modelled

e. Post processing

After analysis of the confusion matrix, confusing pairs are identified. Script specific features can be used to resolve ambiguities in confusing characters.

IV. SURVEY ON OHCR

In the following sub-sections we discuss advancements in the nine Indian scripts. In literature we have seen so far, each researcher used his own database for their experimentation. So it is difficult to compare the various techniques and methods used by the researchers. In the following sections we illustrate the advancements in 7 Indian Scripts:, Assamese, Bangla, Devanagari, Gurumukhi, Kannada, Tamil and Telugu, in terms of preprocessing techniques, feature extraction and recognition methods, along with the data collection devices used. If there is any benchmark database available in a particular script, the database specifications are also discussed.

A. Assamese OHCR

Assamese is the official language of state Assam. It's spoken by nearly twenty million people, including people in some northeastern states. The Assamese script consists of 11 vowels, 41 consonants is given in Figure 4.

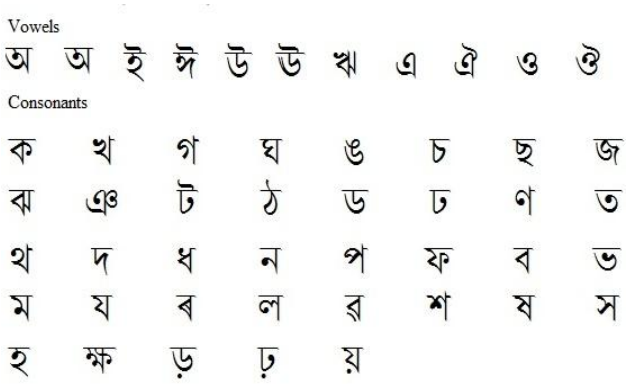

Fig. 4: Assamese Characters

For the recognition of Handwritten Assamese Numerals G. S. Reddy in [30], combine online and offline methods. Online handwritten numeral is recognized using (x, y) coordinates as a feature and Hidden Markov model is used for recognition. Offline handwritten numeral is recognized using features-Vertical projection profile and

horizontal projection profile, zonal discrete cosine transform, chain code histogram, pixel level information and Vector Quantization as recognition model. Ten Assamese numerals total of 207 samples from local Assamese writers is collected using an HP tablet PC. There is a significant rise in performance of combined online and offline system when compared to individual systems. A large database of digits is used by G. S. Reddy in [31] when compared to previous 207.

The data are collected from HP tablet PC. The features are also extended by adding first and second derivatives of the coordinates. In some confusing pairs end distance is also considered as an additional feature. HMM is used as modeling technique and recognition accuracy of 96% is observed. The combined system of HMM and SVM is proposed by Sarma, Bandita in [32]. The recognition accuracy is 98% with the features (x, y) coordinates, first and second derivatives. Performance comparison of HMM and SVM modeling is done in [33] over 147 Akshras (each 1000 samples),which are prepared by the Resource Centre for Indian Language Technology Solutions (RCILTS), Indian Institute of Technology, Guwahati. HMM works well with the features (x, y) coordinates, first and second derivatives, but SVM accuracy declined when we include second derivative.

As there is no benchmark dataset available for Assamese, Udayan in [34] describes the data, which are collected from i-ball 8060 U tablet. The dataset contains a total of 8,235 characters from 183 classes, consists of numerals, basic characters. Combination of different features is used in feature sets and SVM is used as recognition technique. The recognition accuracy of 99.11% is obtained over dataset of 8,235. Accuracy of Akshara recognition is compared with stroke recognition in [35]. The strokes are recognized using HMM and Aksharas are recognized using eight language rules. It is observed that recognition accuracy at Akshara level is low in comparison with stroke as misclassification of the complete Akshara may occur due to single stroke mismatch.

a. Bangla OHCR

Although the whole character set of Bangla is very large, it has only 45 basic characters consisting of 12 shapes corresponding to basic vowels and 33 shapes corresponding to basic consonants. The 50 Bangla basic characters shown in the Fig 5.

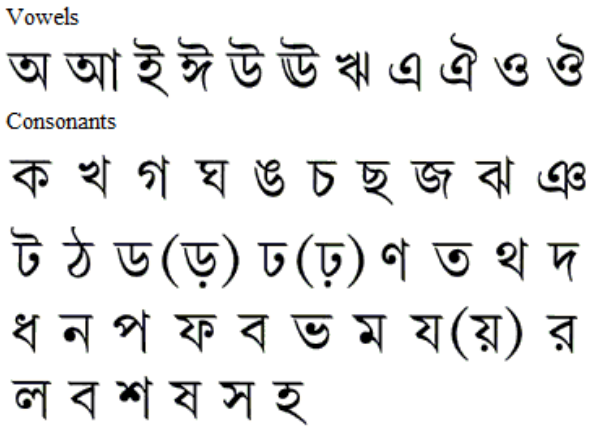

Fig. 5: Bangla Characters

For online Bangla HCR, the features based on Matra are used in [36]. For the classification task Quadratic classifier is used. The dataset of 2500 Bangla numeral data and 12500 Bangla Character is used, which are collected by using mouse and WACOM tablet. The classification accuracy of 98.42% is shown in numeral data and 91.13% in character data. The direction code based feature extraction is used over online basic Bangla handwritten characters in [37]. The sample size is 7043, collected using WACOMIntuous 2tablet from 114 users. Multilayer perceptron (MLP) is used as recognition technique. The recognition accuracy of 93.90% and 83.61% is obtained on its training and test setsrespectively. Two stage character recognition is proposed in [38]. Strokes are classified using HMM. Basing on the predefined lookup table character is recognized in second stage.

The angular feature is applied to describe strokes. The training and testing of the proposed scheme have been done using a database consisting of 24,500. The efforts in the Bangle database creation are described in [39] consist of 25,948 samples of online handwritten Bangla basic characters.

During this stroke are segmented into sub strokes using angle variation. The features for each sub stroke are 8 scalar features representing its shape, size and relative position. The results are compared using classification techniques HMM and nearest neighbor based on DTW. A stroke order free user adaptive approach is proposed in [40]. The system captures as soon the user writes the character and strokes are identified using DTW distance metric. These recognized strokes are used in adaptable Look-Up-Table in character classification.

b. Devanagari OHCR

A Devanagari script has forty-four primary characters, of which eleven are vowels and thirty-three are consonants. The character list is given in the Figure 6.

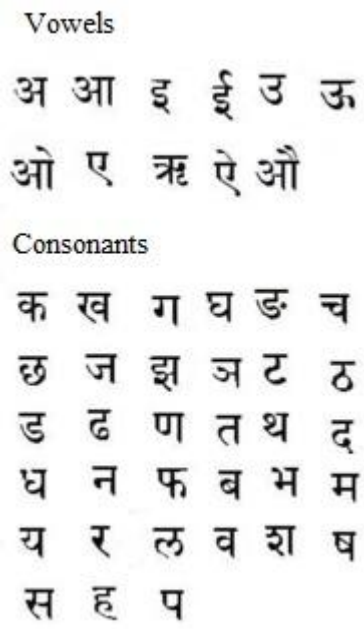

Fig. 6: Devanagari Characters

Recognition in three stages is proposed in [41]. The three stages are Structural Recognition, Feature Based Mapping and Output Mapping. In structural Recognition shirorekha, anuswara, visarga, Chandra Bindu and ardhachandra are detected and removed from the test pattern. After removing special strokes, residual pattern is classified using subspace method. The output of structural decomposition and feature based recognition is mapped to give the final output class label. The features are simply pre-processed (x, y) coordinates. The recognition accuracy of 94% is reported over data collected from 20 writers of 100 frequently used characters, each character collected in seven samples by using HP Tablet PC TC1000. Multiple SVMs are used for stroke recognition in [42].

The features are pre-processed (x, y) coordinates. The strokes are categorized in small and large length stroke basing on average curve length. Small strokes are passed to both SVMs, as most of the time small strokes are given misclassified results. The recognition accuracy of 97% has shown. The system is trained on the data collected from 90 writers by using Super Pen a product of UC Logic. Strokes are represented by spatiostructural features in [43]. The modeling technique SVM is proposed. The recognition accuracy of 95% is obtained. The optimal classifier collection is proposed in [44] by applying genetic algorithm. Set of 25 SVM-based classifiers are constructed using various features and varying SVM kernel parameters. A collection consisting of 5, 10 and 15classifiers are selected from the pool of 25 classifiers for recognition.

The recognition accuracy is 97% for collection of 5 classifiers. Direction based feature is applied in [45]. The feature includes both pen-tip position and tangent angles, which are sampled from the trajectory of the pen-tip, preserving the directional property of the trajectory path. The feature vectors are compared using DTW. Strokes are clustered based on spatial similarity, i.e. their position can be any of the following top-left (T–L), top (T), top-right (T–R), bottom-left (B–L), bottom (B), and bottom-right (B–L). Each character is represented by two features: direction information and positional information in [46]. The feature vectors are compared using Levenshtein Distance Metric. A Framework for On-Line Devanagari Handwritten Character Recognition is illustrated in [47]. The extended directional feature is computed by computing all directions between one critical point and all other critical points. Fuzzy Directional Features are also applied, which is able to include directional variance in the handwritten primitives. Observed results are illustrated that Fuzzy Direction features outperform directional features.

c. Gurmukhi OHCR

The word Gurumukhi has been used term for the Punjabi script. The character set of Gurumukhi shown in the Figure 7.

Fig. 7: Gurumukhi Characters

Pre-processing techniques of Gurumukhi characters are described in [48], Bezier interpolation is applied. The chain code estimation method is applied for slant correction of Gurumukhi character. To store stroke template database, XML has been used. Four attributes have been used in XML database: primary key, (x ,y) coordinate of a point and angle at that point. For recognizing online Gurumukhi characters elastic matching is applied in [49], which does not require a large account of training data and good for writer dependent data. Post processing is applied after elastic matching.

The results are 41% of writers achieve 90% recognition accuracy without post processing, 68% of writers achieve 90%recognition accuracy with post processing. The features during post-processing are all structural features like loops, crossings, headlines, straight lines and dots. A small line segments method is proposed in [50]. The small line segments method is related to chain code. From the input hand written character directions for small line segments are obtained. The directions obtained from an angle range of 30 degrees and forms A to Z 12 segments, then the recognition accuracy with this feature is 95%. The recognition model HMM is applied in [51] and recognition accuracy of 92% is reported.

d. Kannada OHCR

Kannada script has 52 primary characters: 16 vowels and 36 consonants. Each of these can modify a primary consonant to form a compound character or Akshara. The vowels and consonants are given in the Figure 8.

Fig. 8: Kannada Characters

M M Prasad address Divide and Conquer strategy in [52] for Kannada Characters. The compound character is segmented into three separate stroke groups basing on structural and spatial-temporal information. This grouping reduces complexity of huge character combinations. The features (x, y) –coordinates, Trajectory and deviation are used. The modeling technique KNN is used. The recognition accuracy is 81%. Modeling technique Statistical DTW is proposed in [53]. This is the first work in Kannada including all combinations of consonants and vowels, punctuations, Kannada and Indo-Arabic numerals. The comparison with DTW is done, where SDTW has shown enhanced recognition accuracy. The data are collected using a Tablet PC. Article [50] describes the efforts in the MILE database creation by IISC, Banglore. The data are collected from 600 writers to capture variations in writing by using Tablet PC or G-Note. The device captures (x, y) coordinate of the stroke. It also records PEN_DOWN and PEN_UP information. The list includes all the characters, Kannada and Indo-Arabic numerals, punctuations and other symbols.

The output is stored in XML format. The work of online recognition in combination with offline combination is illustrated in [55]. The online features (x, y) coordinates, Pen direction angle, First and Second derivatives of x and y coordinates are addressed. Offline features Directional distance distribution, transition count and projection profiles are used. SVM is used for online and offline classification. Principal Component Analysis is used to resolve ambiguities in offline classification. The recognition accuracy of 92% is reported. Two stage classification for Kannada is proposed in [56]. Based on the output from the primary classifier, best of the available three classifiers is selected for second stage classification. The confusion matrix of primary classifier is used in identifying next stage classifier.

The features are (x, y) coordinates, Quantized slope, dominant points, and Quartile feature. Primary stage uses (x, y) coordinate and PCA with a nearest neighbor classifier. Basing on the output from the first stage, secondary stage selects DTW classifier with suitable features.

e. Tamil OHCR

The Tamil script has 12 vowels, 18 consonants and one special character. They are given in the Figure 9.

Fig. 9: Tamil Characters

Comparison of various Elastic matching algorithms over Tamil characters is performed in [22]. The features (x, y) coordinates, Quantized slope, Dominant Points have applied. Seven distinct recognition schemes based on DTW are proposed. The schemes are implemented using different combination of features. The data are collected using iPAQ tablet PCs from 20 writers 10 samples for each of 156 Tamil characters. Strokes are represented by spatiostructural features in [43]. The modeling technique SVM is reported. The results illustrate that spatiostructural features are more suitable to strokes having larger curve lengths. Application of Regression techniques for OHCR is illustrated in [57].

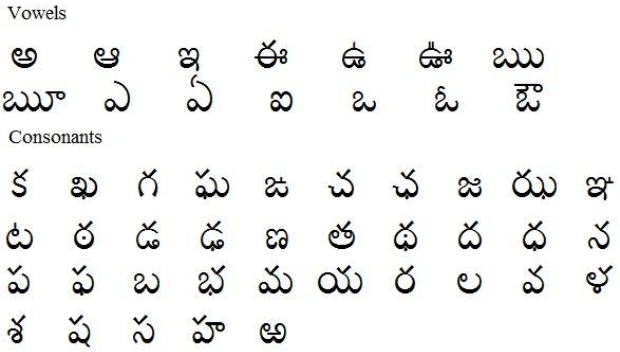

Fig. 10: Telugu Characters

An SVM based stroke recognition method is used in [42] for Telugu characters. Based on proximity analysis, the recognized strokes are mapped onto characters using information of stroke combinations for the script. Each stroke is represented as preprocessed (x, y) coordinates. The data sample of size 37817 is collected from 92 users using the Superpen, a product of UC Logic. The recognition accuracy of 83% is observed. Importance of annotation of online handwritten data is illustrated in [64]. Modular approach for

During this stroke are segmented using velocity profiles or skeletal decomposition. The classifier SVM is used and the features polynomial coefficients and B-Spline coefficients are applied separately. The recognition accuracy, of 68% is observed using polynomial regression and 74% is observed using the B-Spline coefficients.

Dimensionality Reduction Technique PCA is used for Tamil in [58]. In this 15 local features are used in feature vector. The classification techniques two dimensional PCA shows 3% enhancement in accuracy over PCA. The dataset IWFHR 2006Tamil competition dataset is used. The scheme has been used for classification of characters is KNN based on dynamic space wrapping (DSW) of strokes within the characters in [59]. The sub strokes are segmented from stokes using velocity profile. The features Inclination, Proclivity and Curvature are used to describe the stroke.

The recognition accuracy of 92% is reported over dataset collected from 168 users with an average of five trails for each character. Suresh Sundaram has adopted two dimensional PCA and Nearest Neighbor as primary classifier.

Based on confusion matrix analysis, post-processing has been applied over confusing characters and observation is that confused pairs accuracy increase by 2% [60]. The dataset is HP labs IWFHR dataset. Kiran illustrates the advantages of SDTW over HMM in [61] as the best result of HMM can be obtained using simpler models of SDTW. The novel method fractal codes are used for representing characters, which reduces the time while encoding and decoding characters. The classification technique has been proposed is Nearest Neighbor based on DTW. The recognition accuracy of 98% is reported over HP labs IWFHR dataset. The combination of local and global features are proposed in [62] with SVM as the recognition model. The recognition accuracy of 96% is obtained over HP labs IWFHR dataset and 95% is observed over Alpaydin database. A novel algorithm combination Randomized Bayesian network (RBN) classifiers and Expectation Maximization (EM) algorithm is proposed in [63]. In this RBN is trained continuously using EM. This method makes use of expensive unlabeled data and is preferable if labeled data is scarce. The results are illustrated Semi-Supervised Online Bayesian Network Learner has enhanced recognition accuracy over HMM and SVM. The proposed method is evaluated over The IWFHR 2006 competition handwritten online data set.

f. Telugu OHCR

Telugu language found in the South Indian states of Andhra Pradesh and Telangana as well as several other neighboring states. Subset Telugu symbols given in the following Figure 10.

coordinates, normalized first and second derivatives and curvature features are used.

The results of 90% are obtained with the combination of all the seven above mentioned features, over the collected data using Acecad Digital Notepad. A data collection procedure using ACECAD Digimemo is illustrated in [67]. The combination time-domain and frequency domain features is proposed, which enhances performance rather than applying features individually.

V. Scope in Telugu OHCR

In this section we address what are the challenges in Telugu Online HCR, along with the extensions we can apply for Online HCR in Telugu with the survey of other Indian Scripts.

A. Challenges in Telugu OHCR

In this section we address what are the challenges in Telugu Online recognition of strokes is proposed in [65]. Based on the relative position of strokes in the character, the strokes are categorized into baseline, bottom, top strokes. The recognition model SVM is used for each category separately. The recognition accuracy is high for each stage, when compared to combined classifier. Elastic matching technique, DTW is used in [66]. The local features: x-y features, Tangent Angle (TA) and Shape Context (SC) features, Generalized Shape Context (GSC) feature and the fourth set containing (x, y)

HCR, along with the extensions we can apply for Online HCR in Telugu with the survey of other Indian Scripts.

Huge number of Character Classes

In Telugu 13 vowels (V) and 35 consonants (C) are in common usage. A character could be V, C, CV, CCV and also CCCV or numeral. The total possible combinations are listed in the Table 2 below. Recognition time can be reduced, if we reduce the huge character set.

Table 2: Combination of Telugu Characters

Character	Type
V	13
C	35
CV	455
CCV	15925
CCCV	557375
NUMERALS	10
Total	573813

Similarity among many classes of characters

In Telugu script, many of the characters resemble one another in structure. Further, many users write two or more characters in a similar way which can be difficult to classify correctly. The performance of the system may be enhanced by employing classifiers trained on most confusing pairs.

Primitives of Writing

Some of the considerable writing primitives of a stroke are Number, shape and size, speed, direction, curvature etc. Primitives of writing always vary from one user to another and time to time also. There is no guarantee that same user writes in the same way after some time. Manual analysis of the recognition results reveals that a few classification errors are due to variation writing primitives.

Devices Used

Pens's physical position on a tablet is mapped into discrete (x, y) pixels in screen coordinates, due to digitization there may be information loss. The Noise is due to physical condition of devices, for example, dirt on the tablet surfaces. Literature shows that preprocessing of online data results in enhanced accuracy [48].

B. Future Work

More than one decade of research in Online Indian HCR, not much work has been done towards online handwritten character recognition in Telugu. With the review of literature on Telugu OHCR the following extensions we can propose:

Dataset

In a research area related to pattern recognition Benchmarking database is very important. In Telugu the dataset available is Hp-Labs data in UNIPEN format. The data are collected using Acecad Digimemo electronic clipboard devices using the Digimemo-DCT

application. Literature survey shows very less research done using HP-Labs dataset and researchers use their own databases for evaluating their techniques. If the standard database is used, it will be good to compare various techniques have been proposed by the researchers,. To increase accuracy some preprocessing techniques can also be applied over this dataset.

Feature Extraction

Feature extraction plays key role in the stages of character recognition. Performance of recognition depends on features input to it. The study on OHCR illustrates most of the features used so far are local features [13]. These describe the stroke at the point level, so fail to capture global information. Global features try to capture the overall information. Dimensionality reduction techniques can also be applied to reduce feature space. In addition to that we can explore more local features which can discriminate confusing character correctly. A fusion of offline and online features also result in enhanced accuracy of the survey [55]. These techniques can also be applied for Telugu OHCR.

Classification Techniques

Most of the recognized methods have been used in literature are HMM and SVM. The highest recognition accuracy of 90% is reported in literature so far. We can delve into other classification techniques which can increase accuracy further. To obtain computationally more efficient recognition method, classifier schemes can be combined. For optimal classifier combination we can go for soft computing approaches. It is observed that multi-level classification schemes are also popular with increased efficiency.

Post processing

After analysis of the confusion matrix, confusing pairs are identified. Script specific features can be used to resolve ambiguities in confusing characters [60]. Thus, these post processing steps have a good scope to improve the performance of online Telugu handwritten character recognition.

VI. CONCLUSION

In this article a review of data collection tools, datasets, feature extraction techniques and classification models of Indian Scripts for Assamese, Bangla, Devanagari, Gurumukhi, Kannada, Tamil and Telugu are explored. Literature review shows that the Support Vector Machine is the prime choice among researchers next to a Hidden Markov model. Hybrid methods can also be applied in terms of combination of feature extraction techniques to classify the characters properly for training purpose. We also provided what is the future extension we can do in Telugu OHR with the advancements in the remaining Indian scripts. We wish this article not only provides insight of Indian Online HCR, but also provides in depth information for future endeavors. The Resource Centre for Indian Language Technology Solutions (RCILTS) and Technology Development of Indian Languages (TDIL) are established for the purpose of promoting research in Information Processing Tools and Techniques in Indian languages. They are providing financial support for researchers in HCR related fields in India.

REFERENCES

[1] Plamondon, R., Srihari, S.N."Online and Offline Handwriting Recognition: A Comprehensive Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 22(1) (2000) 63-84
[2] Bharath A. and SriganeshMadhvanath,"Online Handwriting Recognition for Indic Scripts", HP Laboratories, India, HPL-2008-45, May 5, 2008
[3] R. M. K. Sinha, "A Journey from Indian Scripts Processing to Indian Language Processing", IEEE Annals of the History of Computing, pp8-31, Jan–Mar 2009.
[4]_____https://www.wacom.com/~/media/files/store-manuals/legacy-models/intuos-users-manual-win.pdf
[5] http://www.iball.co.in/category/Pen-Tablets-/217
[6] http://www.acecad.com.tw/index.php/products/digital-notepad-with memory/digimemo-a402
[7] http://lipitk.sourceforge.net/dataset-tools.htm
[8] Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. Unipen project of on-line data exchange and benchmarks. In 12th Int. Conference on Pattern Recognition, pages 29–33,1994
[9] Y.M. Chee, M. Froumentin, and S. M. Watt. Ink Markup Language (InkML). Technical report W3C, October 2006.
[10] M. Agrawal, K. Bali, S. Madhvanath, and L. Vuurpijl. UPX: A new XML representation for annotated datasets of online handwriting data. In Proc. ICDAR, pages 1161–1165, 2005.
[11] http://lipitk.sourceforge.net/hpl-datasets.htm
[12] http://www.isical.ac.in/~ujjwal/download/database.html
[13] M. Mori, S. Uchida and H. Sakano, "Global feature for online character recognition," Pattern Recognition Letters, 35, 142–148, 2014.
[14] E.J. Bellegarda, J.R. Bellegarda, D. Nahamoo, K.S. Nathan, A probabilistic framework for on-line handwriting recognition, Proceedings of IWFHR III, Bu!alo, New York, May 25}27, 1993, pp. 225-234.
[15] J.M. Hollerback, An oscillation theory of handwriting, Biol. Cybernet. 39 (1981) 139-156.
[16] R. Plamondon, F.J. Maarse, An evaluation of motor models of handwriting, IEEE Trans. Systems Man Cybernet. 19 (5) (1989) 1060-1072.
[17] L.R.B.Schomaker,H.-L. Teulings,A handwriting recognition system based on the properties and architectures of the human motor system, Proceedings of the IWFHR, CENPARMI Concordia, Montreal, 1990, pp. 195-211.
[18] K. H. Aparna, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V. S. Chakravarthy, SriganeshMadhvanath, "Online handwriting recognition for Tamil", Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR' 04), Tokyo, Japan, 2004, pp 438-443.
[19] Uchida, S. and Sakoe, H., "A survey of elastic matching techniques for handwritten character recognition," IEICE Transactions on Information and Systems E88-D(8), 1781–1790 (2005).
[20] K.F.Chan, &D.Y.Yeung, "Elastic structural mapping for online handwritten alphanumeric character recognition," Proc. of 14th International Conference on Pattern Recognition,Brisbane, Australia, August, 1998, pp 1 508-1511.
[21] L. Li, L. Zhang and J. SU, "Handwritten character recognition via direction sring and nearest neighbor matching", The Journal of China Universities of Posts and Telecommunications, vol.19,No.2 (2012) October, pp. 160-165,196.
[22] N. Joshi, G. Sita, A. G. Ramakrishnan, and S. Madhvanath, "Comparison of elastic matching algorithms for online Tamil handwritten character recognition," in International Workshop on Frontiers in Handwriting Recognition, (Tokyo), pp. 444–449,October 26-29 2004.
[23] W. Zeng, X. Meng, C. Yang, L. Huang, Feature extraction for online handwritten characters using Delaunay triangulation, Comput. Graph. UK 30 (2006).
[24] S. J. Cho and J. Kim, "Bayesian network modeling of strokes and their relationships for online handwriting recognition," in International Conference on Document Analysis and Recognition, Seattle, USA, pp. 86–90, September 10-13 2001.
[25] T.M.T. Do and T. Artieres, "Conditional Random Fields ` for Online Handwriting Recognition",th International Workshop on Frontiers in Handwriting Recognition, 2006

[26] D. D. Kerrick and A. C. Bovik, "Microprocessor-based recognition of handprinted characters from a tablet input," Pattern Recognition, vol. 21, no. 5, pp. 525–537, 1988.
[27] E. Poisson, C. V. Gaudin, and P.-M. Lallican, "Multi-modular chitecture based on convolutional neural networks for online handwritten character recognition," in Proceedings of the 9th International Conference Neural Information Processing, vol. 5,(Singapore), pp. 2444–2448, November 18-22,2002
[28] S. Marukatat, T. Artires, B. Dorizzi, and P. Gallinari, "Sentence recognition through hybrid neuro Markovian modelling," in International Conference on Document Analysis and Recognition, Seattle, Washington, USA., pp. 731–735, September 10-13 2001
[29] Graves, S. Fernandez, M. Liwicki, H. Bunke, and J. Schmidhuber, "Unconstrained online handwriting recognition with recurrent neural networks," in Advances in Neural Information Processing Systems
20, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008.
[30] G. S. Reddy, P. Sharma, S. R. M. Prasanna, C. Mahanta, and L. N. Sharma. Combined online and offline assamese handwritten numeral recognizer. in Procedings of 18th National Conference on Communications (NCC-2012), pages 1-4, 2011.
[31] G. S. Reddy, B. Sarma, R. K.Naik, S. R. M. Prasanna and C. Mahanta, "Assamese Online Handwritten Digit Recognition System using Hidden Markov Models", accepted at the Workshop on Document Analyisis& Recognition, 2012.
[32] Sarma, Bandita, et al. "Handwritten Assamese numeral recognizer using HMM & SVM classifiers", National Conference on Communications (NCC), 2013.
[33] D. Das, R. Devi, S. Prasanna, S. Ghosh and K. Naik "Performance comparison of online handwriting recognition system for assamese language based on hmm and svmmodelling", International Journal of
Computer Science & Information Technology, vol. 6, no. 5, 2014 .
[34] UdayanBaruah, Shyamanta, Hazarika," A Dataset of Online Handwritten Assamese Characters", Journal of Information Processing Systems,2014.
[35] S R M Prasanna, Rituparna Devi, Deepjoy Das, SubhankarGhosh, Krishna Naik,"Online Stroke and Akshara Recognition GUI in Assamese Language using Hidden Markov Model", International Journal of Scientific and Research Publications, Vol. 4, Issue 1, Jan 2014.
[36] K. Roy, N. Sharma, T. Pal and U. Pal, "Online Bangla Handwriting Recognition System", In WSPC-Proceedings (ICAPR 2007)
[37] U. Bhattacharya, B. K. Gupta and S. K. Parui, "Direction Code Based Features for Recognition of Online Handwritten Characters of Bangla", Proceedings of 9th ICDAR, pp. 58-62, 2007, IEEE Press.
[38] S. K. Parui, K. Guin, U. Bhattacharya and B. B. Chaudhuri," Online Handwritten Bangla Character Recognition Using HMM, Proceedings of 19th ICPR published by IEEE Computer Society Press,
2008.
[39] T. Mondal, U. Bhattacharya, S. K. Parui, K. Das, V. Roy, Database generation and recognition of online handwritten Bangla characters in the Proceedings of the International Workshop on Multilingual OCR (MOCR), Article No. 9, ACM International Conference Proceeding Series,
Barcelona, Spain , 2009.
[40] D. Dutta, A. Roy Chowdhury, U. Bhattacharya and S. K. Parui, Stroke level User Adaptation for Stroke Order Free Online Handwriting Recognition, Proc. of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014) held at Crete, Greece during 1-4 September 2014, pp. 250 - 255.
[41] N. Joshi, G.Sita, A.G. Ramakrishnan, V. Deepu, S. Madhvanath,, "Machine Recognition of Online Handwritten Devanagari Characters", In Proceedings of the Eighth International conference on Document Analysis and Recognition 2005.
[42] HariharanSwethalakshmi, AnithaJayaraman, V. SrinivasaChakravarthy and C. Chandra Sekhar, "Online Handwritten Character Recognition of Devanagiri and Telugu Characters using Support Vector Machines," Proceedings.of the 10th International Workshop on Frontiers in Handwriting Recognition,October 2006.
[43] H Swethalakshmi, C Chandra Sekhar, V S Chakravarthy, "Spatiostructural Features for Recognition of Online Handwritten Characters in Devanagari and Tamil Scripts, Proceedings International Conference Artificial Neural Networks, Vol 2, pp.230-239, (2007).
[44] Jitendra Kumar, VS. Chakravarthy, "Designing an optimal Classifier Ensemble for online character recognition using Genetic Algorithms", 11th International Conf. on Frontiers in Handwriting Recognition, Montreal, Canada, 19--21 August 2008
[45] Santosh, K. C., Nattee, C., and Lamiroy, B. (2012). Relative positioning of stroke-based clustering: a new approach to online handwritten devanagari character recognition",International Journal of Image & Graphics, 12(2), 1250016.
[46] S. DuttaChowdhury, U. Bhattacharya and S. K. Parui, Online handwriting recognition using Levenshtein distance metric, Proc. of the 12th International Conference on Document pp.79-83,2013. [47] Sunil Kumar Kopparapu, Lajish VL," A Framework for On-Line Devanagari Handwritten Character Recognition"
[48] AnujSharma , R. K. Sharma , Rajesh Kumar, "Online preprocessing of handwritten Gurmukhi Strokes", Machine Graphics & Vision International Journal, v.18 n.1, p.105-120, January 2009
[49] Sharma, A., Sharma, R.K., Kumar, R.," Recognizing Online Handwritten Gurmukhi Characters Using Elastic Matching", In International conference on Image and Signal Processing, IEEE, Los Alamitos (2008)
[50] Anuj Sharma, Rajesh Kumar and R. K. Sharma, "Recognizing Online Handwritten Gurmukhi Characters using Comparison of Small Line Segments", International Journal of Computer Theory and Engineering Volume 1, No. 2, pp.133-137 (2009).
[51] Anuj Sharma, Rajesh Kumar and R. K. Sharma,"HMM based Online Handwritten Gurumukhi Character Recognition"In a Journal of Machine Graphics & Vision International Journal Volume 19 Issue 4, April 2010
[52] M M Prasad, M Sukumar, A G Ramakrishnan, "Divide and conquer technique in online handwritten Kannada character recognition", Proceedings of MOCR (2009),pp.1-6.
[53] K Rituraj, P Mohan, K Shashikiran and A G Ramakrishnan, "Unrestricted Kannada online handwritten akshara recognition using SDTW", Proceedings of ICSPCOM (2010) 1-5.
[54] B. Nethravathi, C. P. Archana, K. Shashikiran, A. G. Ramakrishnan, V. Vijay Kumar, "Creation of a Huge Annotated Database for Tamil and Kannada OHR", Proc. of 12th ICFHR,2010, pp. 415-420.
[55] R. Rampalli, A. G. Ramakrishnan, "Fusion of complementary online and offline strategies for recognition of handwritten Kannada characters," Journal of Universal Computer Science, vol. 17, pp. 81-93, 2011.
[56] Venkatesh Narasimha Murthy and A. G. Ramakrishnan, "Choice of Classifiers in Hierarchical Recognition of Online Handwritten Kannada and Tamil Aksharas," Journal of Universal Computer Science, Vol. 17, pp. 94-106, 2011.
[57] Reddy, N., et al. "Online Character Recognition using Regression Techniques in Applications of Computer Vision ,2008. WACV 2008. IEEE Workshop on. 2008.
[58] S. Sundaram, A. G. Ramakrishnan, "Two Dimensional Principal Component Analysis for Online Tamil Character Recognition", in Proc. of 11th International Conference Frontiers in Handwriting Recognition, pp. 88-94, 2008
[59] AmrikSen, G. Ananthakrishnan, Suresh Sundaram, A. G. Ramakrishnan, "Dynamic Space Warping of Strokes for Recognition of Online Handwritten Characters," IJPRAI 23(5): 925-943, 2009
[60] Suresh Sundaram, A G Ramakrishnan, "An Improved Online Tamil Character Recognition Engine using Post-Processing Methods", 10th International Conference on Document Analysis and Recognition, pp 1216-1220, 2009
[61] S Kiran, K S Prasad, R Kunwar, A G Ramakrishnan, "Comparison of HMM and SDTW for Tamil ndwritten character recognition", Proceedings of SPCOM (2010),pp. 1-4.
[62] G. Ramakrishnan and BhargavaUrala, "Global and local features for recognition of online handwritten numerals and Tamil characters," ACM - Proc. International Workshop on Multilingual OCR, (MOCR 2013), 24 Aug. 2013, Washington DC, USA. [63] RiturajKunwar, apada Pal and Michael Blumenstein, "Semi-Supervised Online Bayesian network Learner for Handwritten Characters Recognition", In Proc. 22nd International Conference on Pattern Recognition (ICPR-2014), Stockholm, Sweden, pp. 3104-3109, 2014.

[64] Anand Kumar, A. Balasubramanian, Anoop M Namboodiri, and C.V. Jawahar." Model-based annotation of online handwritten datasets. In Proceedings of 10th IWFHR, 2006.
[65] Jayaraman, A., Sekhar, C.C., Chakravarthy, V.S." Modular Approach to Recognition of Strokes in Telugu Script", in 9th International Conference on Document Analysis and Recognition (ICDAR-
2007), Curitiba, Brazil (September 2007)
[66] Prasanth, L., Babu, V., Sharma, R., Rao, G. V., and M., D., "Elastic matching of online handwritten tamil and telugu scripts using local features," in [ICDAR '07: Proceedings of the Ninth International
Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2], 1028–1032,IEEE Computer Society, Washington, DC, USA (2007).
[67] V. Babu, L. Prasanth, R. Sharma, G. Rao, and A. Bharath," Hmm-based online handwriting recognition system for telugu symbols. Document Analysis and Recognition, International
Conference on, 1:63–67, 2007.