OFFLINE TYPED GUJARATI CHARACTER RECOGNITION

Manish Kayasth, Dr. Bankim Patel

ABSTRACT

Character recognition is major concern since its inspiration. So, far very limited progress has been made in it, specifically for Indian languages. In this paper authors have presented recognition of offline computer generated and printed Gujarati characters. To identify characters authors used modify version of Hidden Markov Model (HMM) based algorithm. The system is trained and tested for different font size of Gujarati characters.

Keywords: Offline Typed Characters, Character Recognition, Optical Character Recognition (OCR), HMM

1. INTRODUCTION

Language is a way of communication to the opposite person in our day-to-day activities to express our views either in written or oral mode. Prior to IT, important documents are created using language mainly by way of writings on a piece of paper either by handwritten or by typewriter. As a result massive volume of paper documents generated. Further, such documents one has to preserve for long time looking to their purposes, importance, and some time due to their uniqueness. For example, there may be millions of handwritten or typed judgment documents gathered over a period of time in the different courts of India, some historical document like the letters written by Mahatma Gandhi etc. In a certain situations access to such documents is required frequently. So, if those documents are not handled carefully then they may be damaged as well as lost over a period of time. Also access to such documents should be trouble-free. So, because of all these it is necessary to convert them into some other form say in digital form.

Scanning is one of the approaches to convert documents in to digital form. Scanned image needs a lot of preprocessing including filtering, smoothing, slant removing and size normalization before recognition process [14]. The method that is used to convert scanned paper document into identified and editable electronic form is called Character Recognition (CR) technique [12, 13]. It is very important and vital issue now a day [11, 14]. CR is further complicated because of multiple patterns representing a single character, cursive representation of letters, and the number of disconnected and multi–stroke characters. Authors have addressed this complicated subject [11]. In fact, it can be said that CR is still an open problem [11, 14].

The technology use to identify handwritten, typed text is named as Optical Character Recognition (OCR). OCR is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by scanning the paper [12]. The field of OCR has been extensively researched in the past 60 years, and due to its many different applicable environments, it continues to be a rich area for active research [10]. OCR for Indian languages, in general, is more difficult than for European languages because of the large number of vowels, consonants and conjuncts - combination of vowels and consonants. The inflectional and agglutinative nature of Indian languages makes the OCR task quite challenging. Very little work is found in the literature for recognition of Indian language scripts [9, 18]. Relatively more research works are found for Tamil – one of the official script used in the southern part of India, Sri Lanka, Singapore and Malaysia [2, 15, 21]. However, Telugu is the second most popular language in India [9], some work has been reported on the development of OCR systems for Telugu text also [6, 7, 8, 9, 13]. Some work focuses on the recognition of basic Bengali characters i.e. multi-font Bangla character recognition has been attempted [3]. Language models and computational linguistics as it pertains to Indian languages is an area of recent research [16]. To the best of our knowledge, there is no ready and reliable work available on recognition of Gujarati language characters [19]. So, we have thought of having some tool that recognizes Gujarati documents using OCR technology.

2. CHARACTER RECOGNITION

CR can be categorized in two ways – Online, and Offline - as the way of capturing data [22]. In case of Online recognition approach, character is captured as a stream of two points x and y using an appropriate pen position sensor often called a digitizer, rather than as a bitmap [2, 25]. So, it uses the sequential list of coordinates forming the trajectory information of pen movement [21]. The scripts are usually dealt with pen tip traces from pen-down to pen-up positions [22]. Data are captured during the writing process, which makes the information available on the ordering of the strokes [14, 22]. Offline text recognition technologies use optically scanned static images of the paper documents as its input [22, 25] and features are computed only on the basis of a set of object pixels [21]. That is recognition takes place, on an image captured, once the writing process is over [14]. In this paper, authors have used offline character recognition approach.

Many techniques initially designed for character recognition including recent one neural networks. They have been incorporated to analytical methods for recognizing tentative letters or graphemes. The contextual phase is generally

based on dynamic programming and/or Markov chains using algorithms like Edit distance, Viterbi algorithm, etc. Fruitful research has been realized in recent years in the field of analytic recognition with implicit segmentation using various kinds of HMMs [5, 14, 21]. HMMs and Neural network are also being investigated and used for the text recognition [1]. During the last few decades, significant progress has been done towards the development of similar technologies for different scripts [21]. Related review works can be found in [2, 8, 17, 20, 21].

In the present paper, authors proposed Gujarati Character Recognition using modified HMM – called GCRHMM - algorithm for the computer generated typed as well as printed text. For the purposes of this paper, authors have assumed that the segmentation problem of separating each character from its neighbors has been solved, plus input image for recognition is clean and so there is no issue of any kind of preprocessing task like noise filtering and normalizations of scale.

3. INDIAN COMMUNICATION LANGUAGE – AN OVERVIEW

In India, Eighteen different official languages and Ten official scripts are existing [9, 21, 28]. The Constitution of India recognizes twenty-two languages, spoken in different areas of the country [24, 26, 28]. Each language is made up of its fundamental alphabets and grammar. They differ by varying degrees in their visual characteristics but share some important similarities. However, their complexity depends on numerous vowels, consonants, numbers and other characters of language.

Gujarati is a language from the Indo-Aryan family of languages, used by about 50 million people in the western part of India. It is one of the India's most popular languages mainly used in the Gujarat state [23, 24, 26, 28]. Gujarati-script used to write the Gujarati language. The Gujarati alphabet utilizes overall 75 distinct legitimate and recognized shapes, which mainly includes 59 characters and 16 diacritics. Fifty-nine characters are divided into 36 consonants (34 Singular and 2 Compound (not lexically though)) means ornamented sounds, 13 vowels (pure sounds), and 10 numerical digits [4]. Sixteen diacritics are divided into 13 vowel and 3 other characters. The alphabet is ordered by logically grouping the vowels and the consonants based on their pronunciations [28].

4. RECOGNITION USING GCRHMM

The original concepts of HMM was proposed by A. A. Markov. HMM is a finite set of states, each of which is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called

transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are hidden to the outside [27]. The relevance of HMM's was first demonstrated in speech processing and recognition in the late 1980's. Neighbor areas such as signal processing, and handwritten and text recognition have also benefited almost at the same time. Half a decade later, HMM's spread to many other areas such as image processing and computer vision, biosciences etc. Promising results have been obtained from the use of HMM's in several applications in many different areas [28].

There are two different approaches or methods of HMM 1) Discrete and 2) Continuous. The probabilities for each candidate character are calculated. Then, the probabilities are counted to obtain a final best character-list for character recognition. The observations are continuous so authors have used a continuous probability density function, instead of a set of discrete probabilities. Authors have modified HMM algorithm for Gujarati character recognition called as GCRHMM.

Below figure 1 shows the block diagram of the proposed architectural model for Gujarati character recognition. Initially some operational processes like to create template or grid for the character net, create and load data file etc. are performed. After that the image to be recognized is loaded into the system. The different computational steps like boundary checking, character learning, displaying etc. have been performed on the image. Characters matching and recognizing task are then ensuring by GCRHMM algorithm using data store.

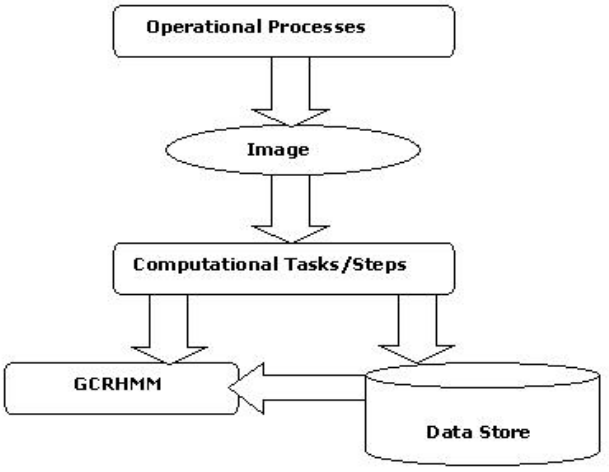

Fig. 1: Architectural model of Gujarati Character Recognition System

Figure 2, given below, shows the components of above shown architectural model. It mainly divides into three components – 1) Learning, 2) Updation, and 3) Identification. In the first part character is created and accordingly system will learn. Different characters' bitmap are stored in the data store which can be further edit and save. Finally, the model uses Neural network based pattern recognition approach for recognizing Offline Gujarati text.

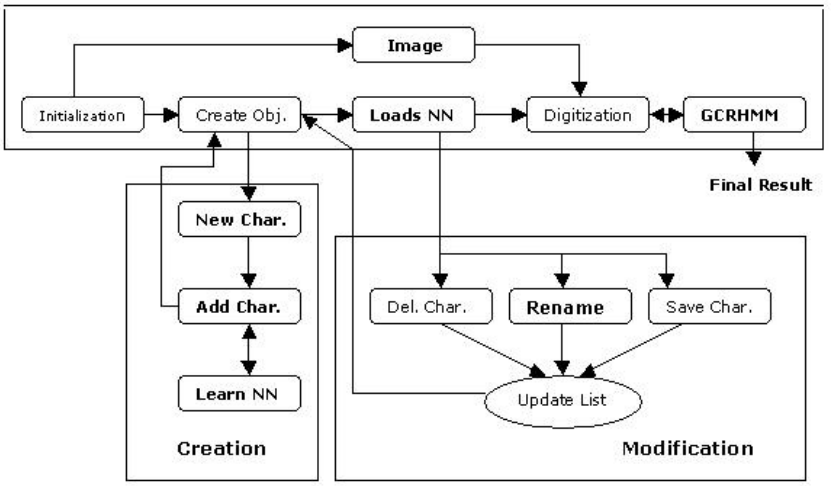

Fig. 2: Core Components of Architecture

Mathematical structure of GCRHMM described as follows:

- The number of states of the model, N.
- The number of observation symbols in the alphabet, M. If the observations are continuous then M is infinite.
- A set of state transition probabilities A = {aij}
- where,

  $a_{ij} = p\{q_{t+1} = j \mid q_t = i\}, \quad 1 <= i,j <= N,$

  where q_t denotes the current state.

Transition probabilities should satisfy the normal stochastic constraints,
$a_{ij} >= 0, \quad 1 <= i,j <= N$ and

N

$\sum_{j=1} a_{ij} = 1, \quad 1 <= i <= N$

A probability distribution in each of the states, $B = \{b_j(k)\}$.

Usually the probability density is approximated by a weighted sum of M Gaussian distributions N,

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N(\mu_{jm}, \sum_{jm}, o_t)$$

where, c_{jm} = weighting coefficients
μ_{jm} = mean vectors
$\sum_{jm}$ = Covariance metrices

c_{jm} should satisfy the stochastic constrains,

$c_{jm} >= 0, \quad 1 <= j <= N, \quad 1 <= m <= M$

and

$$\sum_{m=1}^{M} c_{jm} = 1, \quad 1 <= j <= N$$

- The initial state distribution, $\Pi = \{\Pi_i\}$.
  where, $\Pi_i = p\{q_1 = i\}, \quad 1 <= i <= N$
- Identify number of rows R
- Therefore we can use the compact notation
  $\lambda = (A, C_{jm}, \mu_{jm}, \sum_{jm}, \Pi, R)$
  to denote one with continuous densities.

5. GCRHMM ALGORITHM

The algorithm identifies the characters based on offline identification based on continuous approach. It identifies more than one character. It consists of following steps:

Step 1: Set the [n x m] matrix of grid for the character

Step 2: Create a data store by adding a new character net, delete unwanted character, rename existing character and save the changes of character list into data file.

Step 3: Trains neural-network for all characters pattern.

Step 4: Scans the picture by mapping its surrounding boundaries and converts it into pixel grid. [Digitization]

Step 5: Identify number of rows

Step 6: Check whether other grammatical character is present or not.

Step 7: For Character identification we follow HMM model with its three basic steps, which are as follows:

Step 7.1: Here, we evaluate problem by computing the probability of output sequence of observation p{O|λ} according to HMM.

Step 7.2: After problem evaluation find most likely sequence of states which could have generated given output sequence. An HMM is trained iteratively using GCRHMM algorithm and then used for recognition.

Step 7.3: Once appropriate sequence of state have generated there after it is necessary to find most likely set of state transition and output maximum probabilities p{O|λ}.

Above three steps, compare character's pattern being identify with the entire existing character set's stored in the data store. The character which has the highest probability matching has been displayed as a recognize character.

Step 8: Repeat step-4 to step-7 for the next remaining characters.

6. FEATURE EXTRACTION

The images of all the segmented characters are rescaled into a common height and width producing a grid with say 24 x 32 pixel-size (i.e. shaped-zones) as shown in below given figure 3. The pixel density is calculated as binary patterns and therefore a vector is created. However, due to the varying nature of font-family, there was dissimilarity between the feature vectors of the same class.

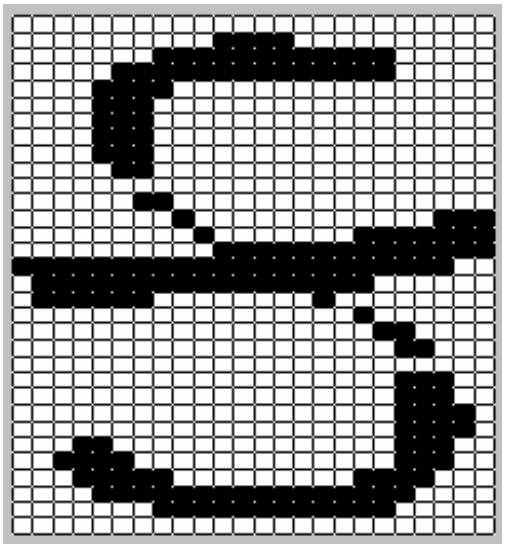

Fig. 3: Creation of machine-generated character's vector

7. EXPERIMENTAL RESULTS

% of Success							
Font	Size	Stored Bitmap – NILKANTH					
		Consonents (34)		Numbers (10)		Vowels (6)	
		72 pts.	**48 pts.**	**72 pts.**	**48 pts.**	**72 pts.**	**48 pts.**
NilKanth	**72**	97.1	100	100	100	83.33	100
	52	100	100	100	100	100	100
	48	100	100	100	100	100	100
	44	100	97.06	100	100	100	100
	40	88.24	91.18	90	90	100	100
	36	44	53	90	90	100	100
	32	58.82	52.94	90	80	83.33	66.67

Table 1: GCRHMM success rate

Authors have used above suggested GCRHMM algorithm to identify Gujarati characters. These experiments use NILKANTH Gujarati fonts with the size of 48 and 72 points as a stored character bitmap which will be created, train, stored and used for the identification purpose. Authors have also used different font-family for these experiments like NIL, NILKANT, GanSyam for the different font-sizes between 24 to 72 points. These experiments covered 34 Consonants, 10 Numbers and 6 vowels. The results of the experiments are shown in the above table. Generated results are fast and highly accurate.

8. CONCLUSION

A continuous GCRHMM is used for the recognition, yielding classification accuracy described by the earlier table as the basis for the NILKANTH fonts with 72 point size and 48 point size. In our system we extract features from a binary pattern. This feature serves to build a set of reference prototypes for the different classes of the character shapes. Recognition is then achieved by simple matching of a candidate character shape to the pre-built prototypes of all the Gujarati Character set. This segmentation free approach proved to be efficient for one font of Gujarati characters. The multi-font aspect is under investigation.

REFERENCES:

1. A. Zidouri, M. Sarfraz – On Optical Character Recognition of Arabic Text - The 6th Saudi Engineering Conference, Vol. 4 – PP 109 – 116, Dec. – 2002
2. A. S. Bhaskarabhatla, S. Madhvanath - Experiences in Collection of Handwriting Data for Online Handwriting Recognition in Indic Scripts - LREC: 4th International Conference on Language Resources & Evaluation, 26-28 May 2004
3. A. Majumdar - Bangla Basic Character Recognition Using Digital Curvelet Transform - Journal of Pattern Recognition Research 1 - 17-26 – 30 Mar. 2007
4. Babu Suthar - Gujarati-English Learner's Dictionary
5. C. Faure & Eric Lecolinet - Survey of the State of the Art in Human Language Technology - OCR: Handwriting
6. C. V. Lakshmi & C. Patvardhan – A High Accuracy OCR System for Printed Telugu text - Conference on Convergent Technologies for Asia-Pacific Region Volume 2, PP 725 - 729 - Issue, 15-17 Oct. 2003
7. C. V. Lakshmi & C. Patvardhan – A Multi-Font OCR System for Printed Telugu Text - Proceedings of the Language Engineering Conference - PP 7 – 2002
8. C. V. Lakshmi & C. Patvardhan - An Optical Character Recognition System for Printed Telugu Text - Pattern Analysis and Applications, Vol. 7, No. 2, PP 190-204 - July, 2004
9. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran – A Bilingual OCR for Hindi - Telugu Documents and its Applications
10. E. Krevat, E. Cuzzillo - Improving Off-line Handwritten Character Recognition with Hidden Markov Models
11. G. Nagy, T. A. Nartker, S. V. Rice – Optical Character Recognition: An Illustrated Guide to the Frontier - Procs. Document Recognition and Retrieval VII, SPIE Vol. 3967, PP 58-69.
12. Guidelines on the Application of New Technology to Population Data Collection and Capture – Chapter 6 – OCR Technology – PP 52 – 62.
13. K. N. Murthy, A. Negi et al – Optical Character Recognition for Oriya Telugu
14. M. F. Zafar, D. Mohamad et al - On-line Handwritten Character Recognition: An Implementation of Counterpropagation Neural Net – Transactions on Engineering, Computing and Technology V10 - Dec. 2005
15. N. J. Rao – Optical Character Recognition for Tamil
16. R. Kasturi & L. O'Gorman et al - Document Image Analysis: A Primer - Sadhana Vol. 27, Part 1, PP. 3–22, Feb. 2002
17. R. M. Suresh, R. J. Kannan et al - Offline Handwritten Tamil Word Recognition Using Hidden Markov Models - 41st Annual National Convention of CSI, Kolkata, Nov. 23 – 25, 2006
18. S. Antani, L. Agnihotri – Gujarati Character Recognition – 5th International Conference on Document Analysis and Recognition PP 418, 1999
19. S. Mehta, S. R. Mohan, J. Dholakia et al - Resource Centre for Indian Language Technology Solutions – Gujarati Achievements

20. S. Hewavitharana & H. C. Fernando et al - Off-line Sinhala Handwriting Recognition using Hidden Markov Models
21. S. K. Parui, U. Bhattacharya, et al – A Hidden Markov Model for Recognition of Online Handwritten Bangla Numerals - 41st Annual National Convention of CSI, Kolkata, Nov. 23 – 25, 2006
22. http://cvit.iiit.ac.in
23. http://ccat.sas.upenn.edu
24. http://languages.iloveindia.com
25. http://cslu.cse.ogi.edu
26. http://india.mapsofindia.com
27. http://jedlik.phy.bme.hu
28. http://en.wikipedia.org