Creation of annotated Tamil handwritten word corpus for OHR

Nethravathi B, Archana C P, Shashikiran K and A G Ramakrishnan

MILE Lab, Department of Electrical Engineering, IISc, Bangalore, India.

{nethra, archana, shashikiran, agr} @mile.ee.iisc.ernet.in

Abstract:

Annotated datasets form a critical aspect in the development of robust technology for handwriting recognition and can be used for comparing results of different techniques used by various research groups. This paper describes the efforts at MILE lab, IISc, to create a database for the design and development of Tamil Online Handwritten Recognition. 100,000 words have been collected from 500 writers in Tamil, so that as much variations in writing style is captured. The data collected incorporated all the symbols (base characters, Indo-Arabic numerals, punctuations and other symbols). An annotation tool has been developed which helps the study of various styles of writing, stroke directions and presence of delayed strokes. Quality tags like class A, B, C etc has been assigned to the words accordingly. The annotated data is stored in a standard XML format defined by OHWR Consortium.

1. Introduction:

Databases are of great importance in any field of research, and handwriting recognition is no exception. A good database of handwritten data can be used to train and evaluate the performance of the recognition engine. Databases for scripts like Roman and Chinese already exist, whereas no such databases exist for Indic scripts. The database collected at MILE lab, IISc contains a comprehensive collection of words in Tamil, collected from many native Tamil people. Predefined word lists have been used to collect data, where the word list covers all the characters in the language. Here the focus is to develop a comprehensive database to support the development of a robust recognition engine. These databases facilitate comparison of different engines and also allow researchers to focus on recognition methodologies. A large database helps in removing bias of the engine towards particular styles of writing.

Tablet PC and G-Note have been used to collect data. The writer writes with an electronic pen on the electrostatic pressure sensitive writing surface of a Tablet PC or G-Note. The device captures the movement of pen tip on its screen in terms of x, y co-ordinates, sampled at equal intervals of time. It also captures the PEN_DOWN and PEN_UP information. The recognition is challenging because of varying styles of writing the same character. This paper describes how the database of 100,000 words has been collected from different schools and colleges, which involved major field work.

The collected data is annotated at the word, stroke group and akshara level using an annotation tool [2] developed by MILE lab. An akshara in Indian languages is a cluster of graphemes that need to be considered together to obtain the correct Unicode representation. Aksharas can be consonants (C), vowels (V) or a combination of them such as CV, CCV and so forth. The output of annotation is stored in the standard XML format [3] which was proposed by the OHWR consortium.

2. OHWR Consortium funded by TDIL:

A consortium made project was funded by Technology Development for Indian Languages (TDIL), Department of Information Technology, Government of India in January 2007 for research on online handwriting recognition. The project aims at developing Online Handwriting Recognition (OHWR) engines for Tamil, Kannada, Malyalam, Telugu, Bangla and Devanagari scripts. We at MILE Lab, IISc, are developing Tamil and Kannada engines. The academic partners of this project are IIT Madras, ISI Kolkata, and IIIT Hyderabad. The private and public industry partners are Learnfun Systems Chennai, CK Technologies Chennai and CDAC Pune.

3. Characteristics of Tamil handwriting:

Tamil compound characters (aksharas) are formed by graphically combining the symbols corresponding to consonants and vowel modifiers using well defined rules. Segmentation of words in these languages is more feasible than it is for English cursive writing as the characters are written separately without much overlap between them. In Tamil script, the majority of vowel modifiers are written as separate symbols and hence they are recognized separately.

4. Selection of complete constituent symbols:

Tamil:

Tamil script comprises 313 characters. Of these, 12 are pure vowels and 23 pure consonants. Thus there are totally 12*23 = 276 consonant-vowel combinations. Apart from these, there are two additional symbols. The set of pure vowels in Tamil and its corresponding transliteration in English is depicted in Fig 1.

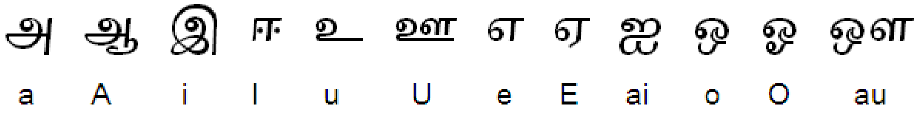

Figure 1. Set of Vowels in Tamil

We have established that only 155 symbols are required to represent all the 313 characters. The details are given below,

1) The vowel modifier for /A/ is depicted by a separate symbol and is written to the right of the consonant. Treating this vowel modifier as a separate class reduces the number of classes. A consonant ட் /T/ combined with the vowel modifiers /a/ and /A/ are shown in two different rows of Fig 2.

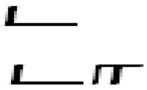

Figure 2. Consonants /Ta/ and /TA/

2) Vowel modifiers of /i/, /I/ and /u/,/U/ create new symbols when combined with the consonants. These new symbols are treated as different classes, thereby adding to the total number of classes. An example of this is shown in Fig 3.

டி டீ டு டூ

Figure 3. Consonants /Ti/, /TI/, /Tu/ and /TU/

3) The vowel modifiers of /e/, /E/, /ai/ are separate symbols written to the left of the consonant. These symbols are also treated as separate classes, further reducing the number of classes. Fig 4 shows an example of a consonant in combination with these vowel modifiers.

டெ டே டை

Figure 4. Consonants /Te/, /TE/ and /Tai/

4) The vowel modifiers of /o/, /O/ have two separate symbols which are written on either side of the consonant. The consonant combined with vowel /o/ will have the modifier of /e/ to its left and the modifier of /A/ to its right. Similarly a consonant combined with vowel /O/ will have the modifier of /E/ to its left and the modifier of /A/ to the right. Since these symbols are already handled separately, the number of classes reduces further. Example of a consonant combined with these vowel modifiers is shown in Fig 5

டொ டோ

Figure 5. Consonants /To/ and /TO/

5) The vowel modifier /au/ also has two symbols with one written on either side of the consonant. The symbol to the left of the consonant is the same as the modifier of /e/ and the symbol to the right is the same as the consonant /La/. These two symbols are already handled separately, similar to case 4, which also causes a reduction in the number of classes. A consonant combined with vowel modifier of /au/ is shown in Fig 6.

டௌ

Figure 6. Consonant /Tau/

Along with the characters, special symbols like full stop and question mark are also incorporated in the symbol list. It is to be noted that in modern Tamil script, Tamil numerals are rarely used. Hence these symbols are not included in our dataset. Hindu-Arabic numerals have been included, and treated as special symbols in our work. The words have been carefully chosen so as to represent all possible symbols used in modern Tamil script.

5. Data Collection for Tamil OHR:

5.1 Criteria for selection of acquisition devices:

The devices used for data collection are the Tablet PC and G-Note. G-Note is more suitable for field work as it is sturdy, affordable and easy to carry. It is also easy for the user to write on a G-Note as the feel is the same as writing on normal paper or pad. The data collected in G-Note is stored as .TOP files. Tablet PC is suitable for individual use. It is heavy and difficult to carry. Also since it is expensive, it cannot be used for field work. The TabletPC data is stored in .txt format. These devices are shown in Fig 7. Predefined word lists have been used to collect data. A Tamil sample handwritten pages is shown in the Fig.8.

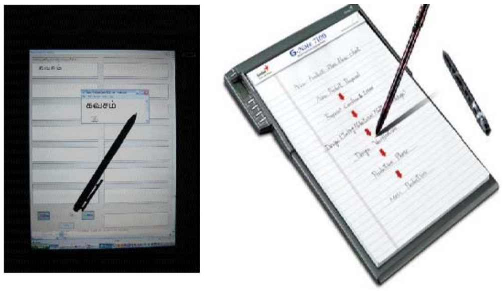

Figure 7. Tablet PC and G-Note 7000

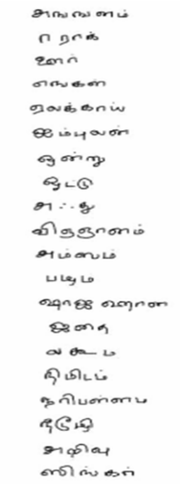

Figure 8. Sample handwritten Tamil Page

5.2 Selection of Writers:

The criteria for selecting writers for data collection was that the person should be a native writer of the language and one who is currently writing regularly. Students and teachers were primarily chosen for data collection as they write regularly.

6. XML Standard for Annotation:

The output of annotation is stored in a standard XML format which has been defined by the OHWR consortium [3]. This standard XML includes all the details about the data, such as the writer details, the device information, the number of pages and words. The words are truthed at the word level annotation. The aksharas and stroke groups are truthed at the character level annotation. All this information is stored in the XML. The XML also contains information about the quality assigned to each word, akshara, stroke group and stroke. This facilitates separation of Class A/good data from the Class R/reject data.

7. Annotation Details:

Once the data is collected, the first process is to do the word level annotation. The collected set has multiple words on each page; hence determined word boundaries are to be used to obtain the strokes of a word. In word level annotation, each word is labeled, using a tool developed by IIIT Hyderabad. The output is stored in a standard XML format defined by the OHWR consortium.

Next is the character level annotation, where the output of word level annotation is given as input. In character level annotation, words are separated into strokes, stroke groups and aksharas and they are labeled. Quality tags are also assigned to them based on the direction of writing, stroke order and validity of strokes. The output at this level is also stored in the standard XML format. This annotation at character level is performed using a tool developed at MILE Lab, IISc [2].

8. Quality labels for Strokes, Stroke groups and Aksharas:

The strokes, stroke groups and aksharas are assigned various quality labels based on the nature of writing. The labels are defined as follows:

Class A: Denotes words written correctly with the expected number of strokes and in the expected direction. They are automatically segmented correctly by the segmentation module. Based on the statistics of writings from a huge number of native writers, there are multiple sets of stroke sequences valid for many stroke groups.

Class B: Denotes words which require manual segmentation and stroke groups with 10% or less overlap.

Class C: Denotes words where two or more normally separate strokes are written as a single stroke or vice-versa. It also includes strokes with overlap greater than 10% and delayed strokes.

Class D: Denotes words with extraneous strokes or overwriting and strokes written in the opposite direction. However, the resulting stroke groups must have the potential to be properly recognized using offline features.

Class R: This is the reject class, containing wrong words and/or strokes for which the likelihood of recognition is very low.

9. Conclusion:

This paper describes how the database has been created for Tamil Online handwriting recognition. The process of creating a reduced symbol list, which includes all the basic symbols of the character set has been described. The focus is on the process of collecting data, the devices used the criteria for selection of writers and why the reduction in number of symbols is required. This paper also tells how the data can be annotated for further use by researchers.

10. Acknowledgment:

This entire data collection effort was funded by Technology Development for Indian Languages (TDIL), Department of Information Technology (DIT), Govt. of India. We thank Mr. Suresh, Mr. Rituraj, Ms. Chandrakala, Ms. Archana, Ms. Saranya and Ms. Southeriya for their efforts which made this task successful. We thank Prof. Deivasundram (University of Madras), AVM Matriculation Higher Secondary School, Govt. Boys Higher Secondary School, Sulur, Presidency College Triplicane, Virugambakkam, Chennai and IIT Madras for contributing to the data set. We also thank Dr. Anoop Namboodiri for giving us the word level annotation tool.

References:

[1] K. H. Aparna, V. Subramanian, M. Kasirajan, G. V. Prakash, and V. S. Chakravarthy. Online Handwriting Recognition for Tamil. Proc. 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9), 2004.

[2] C. P. Archana, K. Shashikiran, and A. G. Ramakrishnan. A Stroke group to Word Annotation Tool for South Indian Languages. submitted to ICFHR 2010, Kolkata, Nov 2010.

[3] S. Behle, S. Chakravarthy, and A. G. Ramakrishnan. XML standard for Indic online handwritten database. Proc. International Workshop on Multilingual OCR, Barcelona, Spain, 2009.

[4] A. S. Bhaskarabhatla and S. Madhvanath. Experiences in Collection of Handwriting Data for Online Handwriting Recognition in Indic Scripts. 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26-28 May 2004.