Creation of a huge annotated database for Tamil and Kannada OHR

Nethravathi B, Archana C P, Shashikiran K and A G Ramakrishnan
MILE Lab, Department of Electrical Engineering, IISc,
Indian Institute of Science (IISc),
Bangalore, India.
{*nethra, archana, shashikiran, agr*} *@mile.ee.iisc.ernet.in*

Vijay Kumar
TDIL Programme, DIT,
Electronics Niketan, 6, CGO Complex
New Delhi.
vkumar@mit.gov.in

***Abstract*—This paper describes the efforts at MILE lab, IISc, to create a 100,000-word database each in Kannada and Tamil for the design and development of Online Handwritten Recognition. It has been collected from over 600 users in order to capture the variations in writing style. We describe features of the scripts and how the number of symbols were reduced to be able to effectively train the data for recognition. The list of words include all the characters, Kannada and Indo-Arabic numerals, punctuations and other symbols. A semi-automated tool for the annotation of data from stroke to word level is used. It segments each word into strokegroups and also acts as a validation mechanism for segmentation. The tool displays the stroke, strokegroups and aksharas of a word and hence can be used to study the various styles of writing, delayed strokes and for assigning quality tags to the words. The tool is currently being used for annotating Tamil and Kannada data. The output is stored in a standard XML format.**

***Keywords*-Annotation; Online character database; OHR database; Tamil handwriting, Kannada handwriting;**

I. Introduction

Databases are of great importance in the field of hand writing research, 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 different samples of words in Tamil and Kannada, collected from different users. 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 engine. These databases allow comparison of different recognition approaches and also free researchers to focus on recognition methodologies. Large databases help in removing the bias of an engine towards specific writing styles.

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, strokegroup and akshara level using annotation tools. 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 [1] which was proposed by the online handwriting recognition (OHWR) consortium.

II. OHWR Consortium funded by TDIL

A consortium mode 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 OHWR engines for Devanagari, Bangla, Kannada, Tamil, Telugu and Malyalam 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.

III. Features of Tamil and Kannada handwriting

Tamil and Kannada conjunct characters (aksharas) are formed by graphically combining the symbols corresponding to consonants, consonant modifiers, and vowel modifiers using well defined rules. Segmentation of words in these languages is more flexible than 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 their models are also built separately. In Kannada, the modifiers are written below or above or beside the base characters which can be considered as separate classes for training and recognition. The script also contains strokes like paadam and dot which can be identified during preprocessing itself.

IV. Selection of complete constituent symbols

A. 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 2 additional symbols. It has been established that only 155 symbols are required to represent all the 313 characters.

The set of pure vowels in Tamil and its corresponding transliteration in English is depicted in Fig 1.

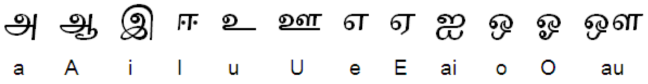


Figure 1. Set of Vowels in Tamil

The symbol reduction procedure is described below [2].

1. The vowel modifier for /A/ is depicted by a separate symbol and is written to the right of the consonant. Treating this 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.

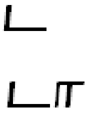


Figure 2. Consonant /T/ modified by the vowels /a/ and /A/

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. Examples of this type are shown in Fig 3.

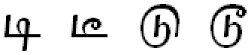


Figure 3. Consonant /T/ in combination with the vowel modifiers of /i/ /I/ /u/ and /U/

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.

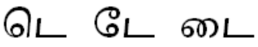


Figure 4. Consonant /T/ in combination with the modifiers of vowels /e/ /E/ and /ai/

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.

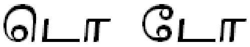


Figure 5. Consonant /T/ in combination with the vowel modifiers of /o/ and /O/

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.

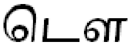


Figure 6. Consonant /T/ in combination with the vowel modifier of /au/

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. HP dataset has been used for training and testing the recognizers. The words have been carefully chosen so as to represent all possible symbols used in modern Tamil script.

B. *Kannada*

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. Thus, there are consonant and vowel modifiers. The total number of possible combinations without reduction in the symbol set is listed in Table I.

Table I
MAXIMUM POSSIBLE COMBINATIONS OF KANNADA BASIC CHARACTERS (V: VOWELS; C: CONSONANTS; N: NUMERALS)

Char Type	V	C	CV	CCV	CCCV	N	Total
Possible Combinations	16	36	576	20736	746496	10	746506

Considering each one of them as a separate class for recognition may reduce the recognition accuracy and increase the computation cost. Also, asking writers to write all the combinations of characters during data collection is not practically feasible. Hence an algorithm has been used to reduce the symbol set.

1. In Kannada, most of the consonant modifiers are written separately from the base characters. These consonant modifiers or *ottus* lie below the base characters, and can be considered as separate classes. The consonant modifiers can be segmented from base character by considering overlap in Y direction. By doing so, the total number of symbols reduces to 645. A few consonants, along with their respective consonant modifiers, are shown in Fig 7.

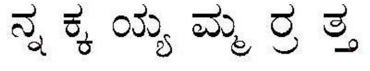


Figure 7. Consonants with their own consonant modifiers.

2. Vowel modifiers are classified into three types in order to further reduce the symbol list. This classification is based on the position, where the modifier is written with respect to the base character.

2a. Some vowel modifiers are written separately below the base character, similar to the consonant modifiers. The same consonant modifier segmentation method can be applied to segment these symbols. This in turn reduces the number of symbols to 543. Examples for this case are shown in Fig 8.

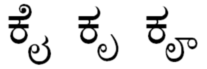


Figure 8. A consonant with the vowel modifiers written below.

2b. Some vowel modifiers are written towards the right side of the base characters as shown in Fig 9. These are also considered as separate classes. Now the number of symbols is reduced to 369.

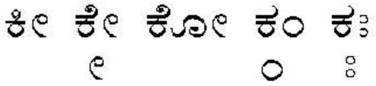


Figure 9. Consonant /k/ with the Vowel modifiers written to its right.

2c. Some vowel modifiers occur to the right of base character with lesser overlap in x direction as shown in Fig 10. Special cases of these modifiers where they are written from below the character are shown in Fig 11. Such cases are taken as separate symbols. This further reduces the number of symbols to 299.

3. Just as vowel modifiers are considered as separate symbols, certain consonants can also be split into different symbols. Examples of such consonants are shown in Fig. 12. Doing so reduces the number of symbols to 291.

4. Some sets of two or more characters differ only by the presence of a dot or dash or both. If those dots and dashes are removed during preprocessing, such characters can be considered as the same symbol. With this, the total symbol list reduces to 251. Such symbols are depicted in Fig 13.

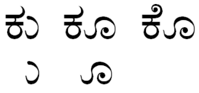


Figure 10. Consonants with the attached vowel modifiers.

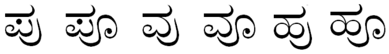


Figure 11. Consonants with vowel modifiers starting from below.

5. Apart from these characters, 9 Kannada numerals, 9 Hindu-Arabic numerals and 22 special symbols which are used in poetry, shlokas and Kannada grammar have been considered.

6. Two more symbols, shown in the top row in Fig 14, are used as consonant modifiers. Also considered are two consonants, which were part of the old character set of the language (see bottom row of Fig 14). So the total symbols become 295. A word list was then created to cover all these symbols.

V. Feasibility of Multiple Segmentation Hypotheses

Since the South Indian languages are never written in cursive style, it is possible to segment them at the level of distinct stroke groups or aksharas. However, occasionally, multiple strokes that make an akshara are combined and written as a single connected stroke. On the other hand, there are people who break a single stroke character or symbol into multiple strokes. These variations can potentially render the segmentation erroneous. However, since, in general, there are only finite ways of breaking a stroke, or combining multiple strokes, it is possible to come up with a strategy that explores alternate segmentation hypotheses and validates one of them after the recognition, based on the confidence level and a verification step. Thus, the segmentation error can be significantly reduced. Such effective segmentation facilitates a recognition engine to successfully deal with unknown proper nouns that normally occur in form-filling applications.

VI. Data Collection for Tamil and Kannada OHR

A. 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 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 are stored as .TOP files. Tablet PC is suitable for individual use. It is heavy

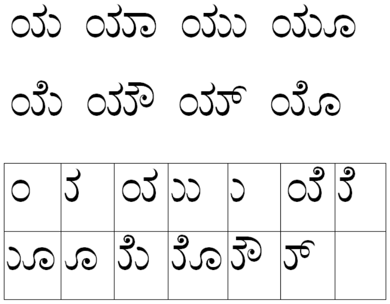


Figure 12. Consonants which can be broken into constituent symbols.

ದ ಧ ಥ

ರ ಠ

ಪ ಫ

ಬಿ ಭಿ ಬ ಭ

ಡ ಢ

Figure 13. Sets of similar Kannada characters with and without dots and dashes.

and difficult to carry. Also since it is expensive, it cannot be used for field work. The data is stored in .txt format. These devices are shown in Fig. 15. Predefined word lists have been used to collect data [4]. Tamil and Kannada sample handwritten pages are shown in Fig. 16.

B. Selection of Writers

The criteria for selecting writers from whom the data was collected was that the person should be a native writer of the language, i.e who is currently writing regularly. Students and teachers were chosen for data collection, since they write regularly.

VII. XML Standard for Annotation

The output of annotation is stored in a standard XML format proposed by the OHWR consortium [1]. 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 strokegroups are truthed at the character annotation level. All this information is stored in the XML. The XML also contains information about the quality assigned to each word, akshara, strokegroup and stroke. This enables in separation of Class A/good data from the Class R/reject data.

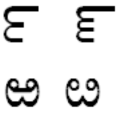


Figure 14. Additional consonant modifiers and old Kannada consonants.

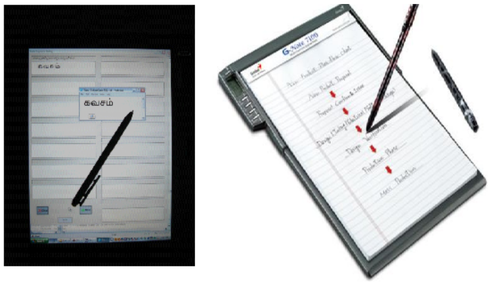


Figure 15. Tablet PC and G-Note

VIII. 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 the 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 we annotate at the character level, where the output of word level annotation is given as input. In character level annotation, words are separated into strokes, strokegroups 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.

IX. Quality labels for Strokes, Stroke groups and Aksharas

The strokes, strokegroups 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 various native writers, there can be multiple sets of stroke sequences valid for any strokegroup.

Class B: Denotes correct words which require manual segmentation and strokegroups 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 but valid strokes.

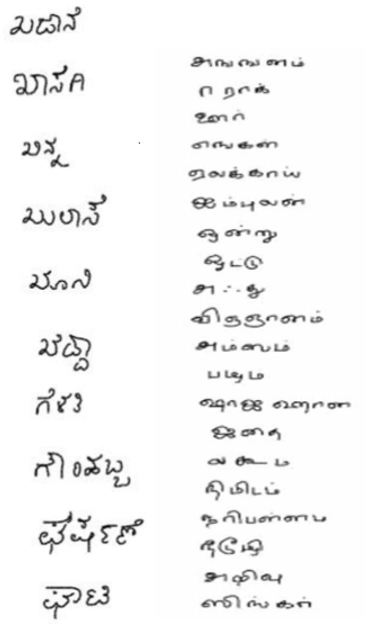


Figure 16. Sample handwritten Kannada and Tamil Page

Class D: Denotes words with extraneous strokes or overwriting and strokes written in the opposite direction. However, the resulting strokegroups must have the potential to be properly recognized using offline features, after removing the extraneous strokes.
Class R: This is the reject class, containing wrong words and/or strokes for which the likelihood of recognition is very low.

X. User-friendly interface of annotation tool

The annotation tool displays the data at three levels -

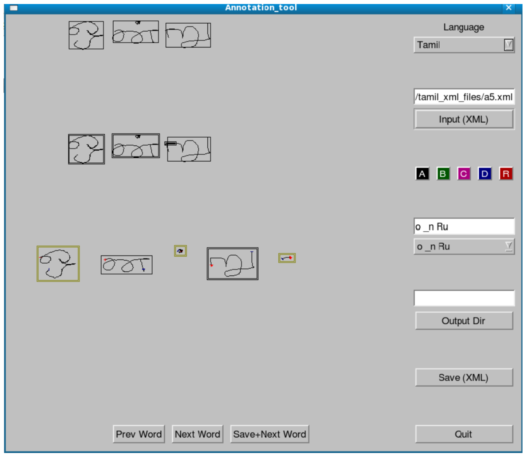
Stroke level: Here, all the strokes of the word are displayed with their order of writing and relative positions preserved. The direction in which each stroke is written is also indicated. This allows the user to assess the quality of the strokes.

Strokegroup level: Here, the strokegroups obtained after segmentation are displayed.

Akshara level: This displays the aksharas, which could be the same as the strokegroups or formed by a combination of the strokegroups.

The strokes and strokegroups such as the ottu and paadam in Kannada and dot in both Tamil and Kannada are indicated in different colors in the display enabling the user to identify segmentation errors, if any.

User interaction with the tool has been kept simple in order to make the task of annotating less tedious. The user can identify wrongly segmented strokegroups/aksharas and quickly modify it by a single click on the corresponding stroke/strokegroup. These clicks automatically update the nodes in the linked list to store the changed information. The quality labels/classes are also provided as buttons, choosing which the user can change the quality tag of a stroke. A change in the assigned quality label of a stroke is indicated by a change in the color of its bounding box in the display. The quality tags of the corresponding strokegroup and akshara get updated automatically. The default and the possible alternate sequence of strokegroup labels are displayed, facilitating the user to choose a different option in the case of a different style of writing. Provision is also given for editing the labels manually. Simple clicks allow the user to save and navigate to the next word or the previous word. All the operations can be performed using the keyboard as well.



Tamil

/tamil_xml_files/a5.xml

o _n Ru

o _n Ru

Figure 17. GUI showing annotation of a Tamil word (IISc OHR, Ver 1.1)

XI. Conclusion

This paper describes how the database has been created for Kannada and Tamil 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. A semi-automated tool for annotating handwritten data from the stroke to the word level is also described. The tool facilitates assigning various quality tags based on the kind of writing and also in checking the effectiveness of the segmentation algorithm.

XII. Acknowledgment

We thank Mr. Rituraj Kunwar, Mr. Suresh Sundaram, Mr. Vijay Mathews, Mr. Chetan, Mrs. SriLakshmi, Mrs. Saraswathi. S, Ms. Chandrakala, Ms. Manjula and Ms. Southeriya for their support in creating this database. We thank Prof. K Narmadamba of Vijaya College, Mr. Manjunath of Maharani Lakshmi Ammani College and Smt.

Vinaya Suvarna of Seshadripuram Girls High school in Bangalore for facilitating Kannada data collection from their students. Our thanks are due to Prof. Deiva Sundaram of Madras University, Presidency College - Chennai, Smt. Lakshmi Balasubramaniam of AVM Meyyappan Matriculation Higher Secondary School in Chennai, Headmaster - Boys Higher Secondary School - Sulur, Headmaster - Girls Higher Secondary School - Sulur, for their benevolent support in collecting Tamil handwritten data from their students. Also special thanks are due to Technology Development for Indian Languages (TDIL), Department of Information Technology(DIT), Ministry of Communication and Information Technology, Government of India for sponsoring this project.

References

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

[2] 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.

[3] 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.

[4] A. Kumar, A. Balasubramanian, A. M. Namboodiri, and C. Jawahar. Model-based annotation of online handwritten datasets. Proc. International Workshop on Frontiers in Handwriting Recognition (IWFHR06), La Baule, Centre de Congreee Atlantia, France, October 2006.