Experiences in Collection of Handwriting Data for Online Handwriting Recognition in Indic Scripts

Ajay S Bhaskarabhatla and Sriganesh Madhvanath

Hewlett-Packard Labs
Bangalore, India
{ajay.b, srig}@hp.com

Abstract

In this paper, we describe initial efforts at Hewlett-Packard Labs, Bangalore, to create datasets of online handwriting in Indic scripts to support research in online handwriting recognition for the Indic scripts. The term "online" here refers to the fact that handwriting is captured as a stream of (x,y) points using an appropriate pen position sensor (often called a digitizer), rather than as a bitmap (image). The paper describes the structure of Indic scripts in brief. It identifies different choices for segmenting characters into simpler shapes that can then be recognized using pattern recognition techniques. The paper discusses these issues in the context of the Tamil script. The remainder of the paper provides an overview of two distinct data collection efforts for the Tamil script - one at the isolated character level, and the other for isolated words. In the context of these efforts, we briefly describe the data collection procedure, tools for collection and subsequent annotation, user-interface issues, the annotation scheme, and the organization of the dataset. The paper concludes with the current status of the effort and future directions.

1. Introduction

India plays host to 18 official languages and 10 official scripts - most of which have not seen much targeted research in human language technologies, despite the large numbers of users. IT penetration is very low (around 2%). Over the years a number of keyboard layouts have been devised for text entry in the different Indic languages, but they remain non-standard and difficult to learn and use, owing to the relatively large number of characters in these scripts. In this setting, technology for the online recognition of handwriting (HWR) has many potential applications - including the creation of appropriate multimodal computing interfaces incorporating the use of speech and handwriting in order to extend the reach of IT to the common man.

One of the major stumbling blocks for language technology research in the Indian context has been the lack of significant shared linguistic resources, and this is especially true for HWR. It is imperative that tools and data formats be standardized and validated datasets be created and made available to change the status quo, and in this paper we describe our first steps in collecting data for Online HWR that can support our own research as well as benefit the research community. Here "online" refers to the fact that handwriting is captured as a stream of (x,y) points using an appropriate pen position sensor (often called a digitizer), rather than as a bitmap (image). It should be mentioned that such datasets would also benefit research in handwritten document analysis, writer identification, script identification, handwritten document indexing and retrieval, and so forth.

There were several options available for data collection. Datasets can be created with focus on a specific application. Datasets can also be created with no specific knowledge of contextual use. Such datasets focus on achieving complete coverage of symbols in the script along with their variations. These 'balanced' datasets are useful for training and evaluation of recognition schemes and provide a common ground for reporting results. Datasets can also be created from archived online handwriting data generated by the use of pen based applications, for example, an "ink chat" application. These datasets may not be complete in terms of the symbols in the script and their variations, but they can be generated quickly and can be used as authentic test data.

Given the dearth of basic linguistic resources for HWR, our initial efforts have been directed towards collection of designed data to support HWR of isolated character and isolated words. In addition to being useful by themselves for applications such as forms automation, these are essential capabilities for the interpretation of larger units of continuous writing.

The rest of the paper is organized as follows. In the next section, we describe some of the salient features of writing in Indic scripts and the challenges they present for recognition. We then describe our efforts at data collection for Tamil at the character and word levels, and briefly discuss the data collection procedure, tools for collection and subsequent annotation, related user-interface issues, the annotation scheme, and the organization of the resulting dataset in the context of these efforts. The concluding section presents current status and future directions for handwriting data collection for Indic scripts.

2. Structure of writing in Indic scripts

The 10 official Indic scripts - Devanagari, Tamil, Gurmukhi, Telugu, Kannada, Gujarati, Oriya, Bengali, Malayalam and Urdu - differ by varying degrees in their visual characteristics, but share some important similarities. With the exception of the Urdu script, they have evolved from a single source, the Brahmi script, first documented extensively in the edicts of Emperor Asoka of the third century BCE. They are defined as "syllabic alphabets" in that the unit of encoding is a syllable, however the corresponding graphic units show distinctive internal structure and a constituent set of graphemes. The formative principles behind them may be summarized as follows (Coulmas, 1999):

- graphemes for independent (initial) Vs
- C graphemes with inherent neutral vowel *a*
- V indication in non-initial position by means of *mātrās* (V diacritics)
- ligatures for C clusters
- muting of inherent V by means of a special diacritic called *virāmā*

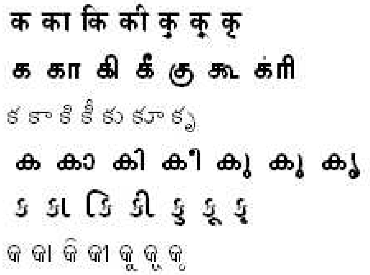

Figure 1: Diversity of Indic scripts

From the standpoint of HWR, an approach based on treating the syllabic units directly as pattern classes has to deal with their large numbers. Most of the Indic scripts have the order of 600 CV units and as many as 20,000 CCV ones in theory, although only a much smaller subset (especially of CCV units) is used in practice. The V diacritics and ligatures for C clusters are not standardized in some scripts. Since handwriting, in the online scenario, is captured as a sequence of pen strokes in writing order the use of larger units also increases the variability in stroke order and hence the intra-class variability for the recognizer.

Approaches based on segmenting syllabic units into the constituent graphemes have to deal with the structural complexity of these syllabic units. In the online scenario, the beginnings of most graphemes are marked by pen-lifts, but not always. In particular, certain V diacritics may be fused inseparably with the underlying C grapheme. Different V diacritics may be visually similar and differ only in how they attach to the C grapheme. Similarly, many of the ligatures for C clusters are non transparent and have to be treated as separate graphemes.

In practice, the approach adopted for HWR is motivated more by pragmatic considerations such as the ease of segmentation of the handwritten word into a smaller number of graphically simpler sub-units, rather than by purely linguistic criteria, and lingusitic interpretation of the recognized units is often relegated to a subsequent stage of processing. As a result, different researchers choose different sets of symbols as sub-word level units for recognition. This is unlike pure alphabetic scripts such as Roman where the choice of symbol set from the perspective of HWR generally coincides with the alphabet for the script. Partly in response to stroke order issues and partly to provide real-time recognition response to pen input, some systems even use individual pen strokes as the most basic set of symbols.

Ideally datasets created to support handwriting recognition should accommodate different choices of symbol sets; however it is not practical to accommodate these in a single annotation hierarchy. One solution is to support several sets of annotation each with its own hierarchy. These hierarchies would be common at the upper levels such as words and syllabic units and diverge thereafter to include different interpretations of symbols and where appropriate, individual strokes.

3. Data collection for isolated Tamil symbols

The present-day Tamil script is simpler than other Indic scripts because of the use of the lack of separate graphemes for voiced, voiceless and aspirated Cs and the vowel muting to unravel C clusters into linear sequences of C graphemes. In addition, some of the vowel diacritics are written in Tamil as distinct symbols to the left and/or right of the C grapheme. This results in Tamil being written as linera as a sequence of visually discrete symbols, which we will refer to as characters, for lack a of better term. However as mentioned earlier, these characters do not have a consistent linguistic interpretation. In addition to independent V and C graphemes, the set includes CV combinations where the vowel diacritics attach above or below the base C grapheme or are otherwise difficult to segment, and those vowel diacritics that occur as distinct characters to the left or right of the base C. The set also includes selected C cluster ligatures and their CV combinations, for a total of 156 charactersFigure 2 shows a Tamil word split into characters.

கால்வலி

க ா ல் வ லி

Figure 2: Tamil word segmented into characters

Like all Indic scripts, there is no tradition of writing Tamil in boxes; however informal observation of several native Tamil writers revealed that they could write characters in boxes consistently with no or minimal training. From the perspective of HWR, being able to “box” Tamil simplifies the segmentation of writing and leads to improved accuracy. Our other motivation to study “boxed” Tamil was to explore generic algorithms for shape recognition that could be used for any choice of symbol shapes from any script. Script-independence of the HWR algorithms was an important design criterion given the large number of scripts to be considered.

In order to build an isolated character recognizer, we collected ten samples of each character from native Tamil writers. The objective of collecting ten samples of each character from each writer was to experiment with writer-adaptation algorithms and compute writer-dependent recognition accuracies (where the recognition engine is trained and tested on samples from the same writer) separately from writer-independent accuracy.

Different kinds of pen-input technology were explored for the purposes of data collection. Initially an application on a PDA (an HP iPAQ PocketPC) was used to present one symbol at a time to the writer, and allow the user to write the symbol once in a set of six boxes that were reused in round-robin fashion. This process was repeated ten times in order to collect the ten samples of each character. This interface is shown in Figure 3.

The data was annotated only at the character level, and the annotation was fully automatic since the writer was expected to write the character presented. The data was manually validated and noisy or invalid data was discarded, and additional samples were obtained from the writer to replace the discarded data.

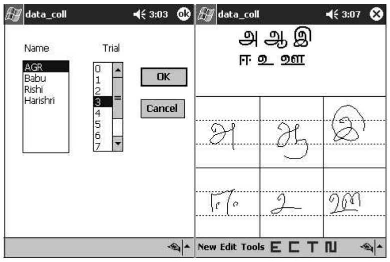

Figure 3: Data Collection using PDA

One of the issues we faced with this interface was the touchscreen (sometimes called a passive digizer) used by the PDA to collect pen input. Casual contact with the screen such as inadvertent resting of the palm or a wayward finger resulted in "ink". This led to some frustration on the part of the writer as well as a higher percentage of invalid data than expected.

A second technology explored was HP Digital Pen and Paper based on Anoto technology (Hewlett-Packard Corp., 2003). This technology uses "digital paper" (paper with a fine pattern of dots printed on it) and a special optical pen which in addition to writing, senses the dot pattern under the pen tip and from it determines its absolute location on the page and stores it as digital ink. Writers were provided with a form with boxes printed on the digital paper (Figure 4) and prompted to write Tamil characters, one in each box. Subsequently, the digital ink stored in the pen was extracted and processed. Again, annotation is automated as the positions of the boxes and the sequence of the tamil characters being transcribed is known. This setup has the advantage of providing a natural pen and paper interface for writing.

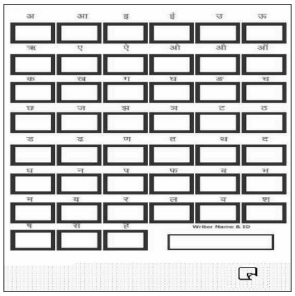

Figure 4: Data Collection using Digital Pen and Paper

The final technology looked at was the TabletPC (HP TC1000). Because of the larger size of the display, the capture application could use large boxes for writing in compared to the PDA. The active digitizer used by the TabletPC overcame the issues of spurious ink encountered with the PDA. The TabletPC also gave the best spatial and temporal resolution (and hence the maximum detail) of the three technologies tried (more than 400dpi, and around 130 samples per second). However it is an expensive option and less suitable for data collection in the field because of much shorter battery life compared to other options.

The ink data corresponding to each symbol is stored in a separate ASCII file tagged by the trial number, and the files are organized into directories by writer ID and symbol ID. Within each file, the handwriting data is stored as a sequence of (x,y) points punctuated by timestamped pen-up and pen-down events.

4. Data collection for isolated Tamil words

The written word is a fundamental unit in any system of writing. From a recognition perspective, the word is especially important in Indic scripts given that sub-word units are a matter of choice and the absence of a tradition of writing words in boxes. The ability to recognize words written continuously (i.e., without boxes) is therefore important for even the most constrained applications such as form-filling.

Most recognition systems attempt to isolate words from larger units of writing using spatial separation criteria, and then use an entirely different set of algorithms to segment them into sub-word symbols and come up with an interpretation of the word based on recognition of the symbols. Our focus here was to collect handwriting data to support research into the recogniton of words once they have been isolated by other means. Our data collection process involved the following steps:

- Identification of symbol set (sub-word units)
- Design of capture text
- Data capture
- Annotation

These steps are addressed in the following subsections.

4.1. Identification of symbol set

In the absence of explicit segmentation provided by boxes, a good set of symbols (sub-word units) is one that balances ease of segmentation of the word into those symbols, with the stability of the symbol pattern across writers (which translates into accuracy of recognition of the symbols). Since Indic scripts do not have a prominent cursive style and pen-lifts may be expected between graphemes, we adopted as the symbol set, the basic graphemes in the script (independent Vs, Cs, V diacritics, vowel-muting diacritic) and added some symbols corresponding to CVs which could not be easily segmented into the constituent base C and V diacritic.

In the specific case of Tamil, this led to the set of 95 symbols that are a subset of 156 characters (Figure 5).

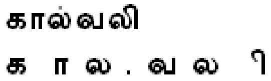

Figure 5: Tamil word segmented in to symbols

4.2. Design of capture text

Owing to the relatively large number of symbols to be recognized, we decided to use an automatic technique for designing the text. First, unique words were extracted from a large Tamil text corpus, and rare and unfamiliar words were discarded based on their low frequency of occurrence. Next, a set cover algorithm was used to extract a minimal subset of words that covered all of the symbols to be recognized. Details of this algorithm have been omitted here for brevity. For Tamil, a set of 60 words was obtained.

4.3. Data capture

Five samples of each word were collected from each contributing writer in five independent trials using a TabletPC application similar to the one used for isolated symbols (Figure 6). In addition to the handwriting data, details collected about each writer included name, age, gender, educational background, and left/right handedness.

As with isolated symbols, the ink data corresponding to each word is stored in a separate file identified by the trial ID. The files are organized into directories. The format of the file is similar to that described earlier for symbols.

4.3.1. Annotation

Unlike the isolated character data collected earlier, word level data has to be explicitly annotated at different hierarchal levels, the final level corresponding to the symbols used by the specific recognition scheme. Annotation at intermediate levels such as syllabic units may also be useful for subsequent linguistic processing. This is an area requiring consensus and standardization, and one possiblity as indicated earlier is to standardize the upper levels of the hierarchy and leave the lower ones (symbol, stroke) to individual researchers.

We have developed a desktop application that can be used for annotation of ink corresponding to words or short phrases at different hierarchical levels, where the hierarchy itself can be defined by the user. The tool allows the assignment of a text or numeric label to a selected set of strokes. In the Indic context, standardization of the names of these labels is also needed.

The process of annotation can be partially automated by bootstrapping word recognition from an initial set of manually annotated word samples, and thereafter using it to suggest segmentation and recognition hypotheses. Further, annotation of one sample can be propagated across the other samples of the same writer who is likely to have similar style of writing across sessions and samples. These ideas are incorporated into our annotation tool (Figure 7).

The annotation is currently stored as part of the ink data file. At each level of the hierarchy, units are denoted by the indices of the constituent strokes, their truth and level of hierarchy they belong to.

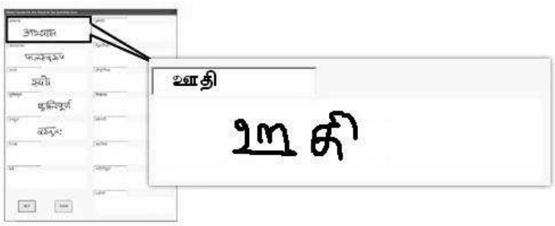

Figure 6: Data capture application UI on Tablet PC

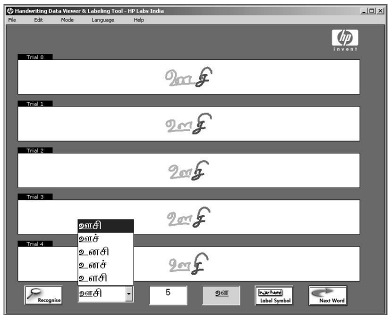

Figure 7: Annotation Tool

5. Future directions

In this paper we have described some of the efforts underway at HP Labs to collect online handwriting data in Indic scripts. Although the specific efforts described are for the Tamil script, the issues and approaches described in the paper are broadly applicable to all Indic scripts which are structurally very similar although graphically different.

The data collection efforts described are a work in progress, and the methodology is still evolving. So far the selection of writers for collecting writing samples has been opportunistic, but it is clearly important to target broad coverage across various parameters like age, handedness, skill, region and style for creating a balanced dataset. Especially for scripts such as Devanagari and Tamil that are used to represent multiple languages and/or used in different geographies, regional differences are very important to capture.

Another important issue is that of standardization of annotation hierarchy while supporting different choices for sub-word units, and the related issue of standardization of labels used at each level of the hierarchy. We are also working on an XML representation for annotation based on the emerging Digital Ink Markup Language standard from W3C (Bhaskarabhatla and Madhvanath, 2004).

Although interactive devices such as the TabletPC and PDAs appear to be sufficient for data collection, they fail to recreate the feel of writing on paper, and people seem to write generally larger on their smoother surfaces. The consequences for handwriting recognition accuracies need to be studied further. Longer term, these efforts need to be scaled to address general handwriting documents, many hundreds of writers, and scripts other than Indic that do not currently have significant linguistic resources.

6. References

Bhaskarabhatla, Ajay S. and Sriganesh Madhvanath, 2004. An XML Representation for Annotated Handwriting Datasets for Online Handwriting Recognition. In *Proc. 4th Int'l. Conf. on Language Resources and Evaluation*. Lisbon, Portugal.

Coulmas, Florian, 1999. *The Encyclopedia of Writing Systems*. Blackwell Publishing.

Hewlett-Packard Corp., 2003. *HP Forms Automation Systems (FAS)*. http://www.hp.com/go/fas.