CURRENT SCIENCE

Volume 107 Number 2 25 July 2014

GUEST EDITORIAL

The magic of automated recognition of handwriting

Handwriting recognition has a variety of applications. Worldwide, the money spent for data entry from handwritten notes, forms and records runs into trillions of dollars. Looking at our 2010 census, the data collection by about 60,000 enumerators in handwritten forms took six months, whereas keying in the data into servers took over two years. In these and other applications mandating handwritten reports such as healthcare and industrial quality control and testing, an instant digital conversion to text will reduce huge costs as well increase productivity.

All Government application forms can be filled using handwriting and can be directly entered into a structured database, if handwriting recognition is perfected, standardized and made ubiquitously available in all computing devices. The additional advantage of such systems is that they can be designed to be user-adaptive, so that they learn the idiosyncrasies of the particular writer to give high performance. Then, it will be better to use handwriting input, rather than keyboard-based input. Interestingly, it is also possible that the user adapts to the system in terms of slightly modifying the shape, direction or order of strokes, when one finds that one's unusual style of writing is not properly recognized by the engine!

Handwriting recognition requires a fusion of intelligent signal processing to extract the most appropriate features, machine learning, natural language processing, incorporating domain knowledge and contextual prediction. It could deal with isolated characters or cursive handwriting, limited or open vocabulary, depending upon the application scenario. Handwriting recognition could be offline or online. Offline recognition deals with scanned images of material that has already been written and is available on paper, palm leaves or other writable materials or surfaces such as rock engravings. This is still an unsolved problem even for Latin script, except for form-filling applications, where isolated characters are written in individual boxes, obviating the problem of segmentation. In online handwriting, the data is captured as the writing proceeds on a special pressure- or touch-sensitive device, so that dynamic information such as the number of strokes, direction of the strokes and velocity are available for every character or word. This availability of temporal information has the potential to increase the recognition accuracy, though unusual writing not conforming to one of the expected directions of writing the strokes is likely to be misrecognized. In such situations, offline recognition might score as long as the final shape of the character is acceptable, since the image captured is independent of the direction of writing and the number and order of the strokes used to write a character.

One could also develop an automated music symbol notation reader, so that a composer can directly write his composition with all the notations and it can get converted to a standard format for display or printing for a book or for his symphony group. Vedic Sanskrit is similar, with symbols for *udata*, *anudata*, *svaritha*, *deergha svaritha* and *plutha*, and an engine that recognizes handwritten *Grantha* or Devanagari with all these symbols will be highly useful. There are organizations which believe that it is possible to predict one's personality from one's handwriting and train people accordingly. Signature verification continues to be the key means of validating cheques by banks. Other useful applications of writer identification technology are in biometrics and forensics. When there are disputes as to whether certain parts of an ancient manuscript were actually written by the author claimed, one can invoke verification based on specific features of someone's handwriting.

Handwriting is a realistic alternative to keyboard input for any computing device. Though QWERTY keyboard is widely used today for all (small alphabet) European languages, and its mapped version for other languages, it is not a convenient interface for languages such as Chinese and Japanese, and even for Indic scripts such as Devanagari, which contain innumerable graphemes born out of consonant clusters. On the other hand, a handwriting input lends itself for inputting such complex scripts and the user also feels that it is a natural interface. Any keyboard has a finite number of combinations or codes possible, whereas handwriting allows limitless variety of input symbols, including sketches, drawing, block diagrams and bullets; editing and annotating, which require heavy interaction, direct pointing and manipulation. The freedom to directly write whatever comes to the mind leaves one to think freely, unlike the case where one needs to invoke multiple tools and menus to input various types of symbols and notations, apart from the text itself. Such an interface is also elegantly user-friendly, since the learning curve is limited, unlike the case of a word processor, where one needs to separately learn how to invoke the equation editor, create block diagrams, tree structures, molecular structures, etc.

Imagine the joy that would be experienced by an academician (especially a mathematician or chemist), if he/she can put in the equations by simple handwriting on a tablet device and prepare his/her journal paper, plenary talk, book or other written material. This is what handwriting recognition is ultimately expected to accomplish at some point.

Authors of books, correspondents of newspapers and magazines, writers in police and lawyers will find it a boon, if they can directly write in the language required and get it recognized and formatted to their requirement. There are other applications of handwriting recognition such as automation of form processing, mail sorting, cheque reading, digital access of manuscripts and self-learning packages for new languages.

Active work on recognition of handwriting in Indic scripts is recent. Researchers in India mainly focused on English and working on Indic languages was not considered fashionable until Technology Development in Indian Languages (TDIL), Department of Information Technology, under the Ministry of Communication and Information Technology started funding research consortia on several language technologies in the country. Under the aegis of TDIL, there is a consortium for Online Handwriting Recognition in Indian languages and currently researchers have developed word-level recognition engines with various levels of accuracy for Tamil, Kannada, Hindi, Malayalam, Telugu and Bangla, and character-level engines for Assamese and Punjabi.

Handwriting can be recognized holistically at the word level or by recognizing the individual characters or symbols. Most Indian languages are not cursively written and hence lend themselves for the approach of individual symbol recognition and then forming the word. Cursive words in English or Bangla are difficult to segment at the character level and hence are better handled by hidden Markov model-based methods with vocabularies. However, languages such as Tamil, Kannada and Telugu are morphologically rich, agglutinative and hence cannot be captured by any length of a dictionary or vocabulary. For example, every Tamil verb root (e.g. /vA/ or /pO/) can give rise to a few thousand (or more) morphologically derived, unique words. Thus, reliable recognition methodologies for such languages cannot use a vocabulary-based selection or correction approach. The right alternative would be to explicitly segment the individual symbols and recognize them using a statistical classifier such as support vector machines, use statistics of co-occurrence of symbols or *aksharas* and expert classifiers at the second level to disambiguate between similar looking symbols, to improve the recognition results at the word level.

In the case of Indic scripts, we need standardization in terms of how to write isolated symbols, characters or syllables within separate boxes. This important problem has not even been addressed by researchers or the Government bodies and this is one of the reasons for lack of progress in offline handwriting recognition in Indic scripts even for form processing. Development and standardization of this technology shall facilitate automated search of a lot of material, for cross-lingual search, verification for employment or forensic purposes.

With mature handwriting recognition technology, it is possible for the common man to write and send SMS in his mother tongue. To facilitate this, the Union and State Governments must compel all the mobile manufacturers and service providers to support Unicode fonts for display and transmission of Indic scripts.

Speech-based input is an equally preferable, if not a better choice. However, current speech recognition technologies require extraordinary amounts of transcripted speech for training, which makes it not at all scalable to new languages. Further, speech recognition is highly susceptible to surrounding noise, whereas handwriting input can be virtually free from noise in outdoor situations. There are lakhs of handwritten ancient scripts on palm leaves in Nagari, Nandinagari and Sharada. There are countable scholars today, who can read such scripts. Unless we take immediate steps to digitize these palm-leaf manuscripts, there is the definite danger of permanently losing them, as well as the knowledge embedded in them. It is impossible to develop high-performance recognition engines for all such handwritten scripts (manuscripts) in anyone's lifetime. Even for English, there is no commercial engine for offline handwritten text. So, a meaningful and practicable way of digitizing such ancient Indic manuscripts quickly is to get the expert read such a document and a scribe to write the same in some modern Indic script and capture it online, so that we can recognize it reasonably well and then render the Unicode version of the document in any Indic script, old or new, at the click of a button. To my knowledge, this is the only viable means of digitizing precious manuscripts.

We need to develop simple handheld devices with handwriting recognition and text-to-speech conversion capabilities in-built, so that persons with speech disability can easily communicate with others by simply writing and getting the device to recognize and read the same.

Development of language technologies such as handwriting recognition, multilingual speech recognition based on breakthrough phoneme-level segmentation and recognition technology and machine translation will go a long way in ensuring that Indic languages will survive in the internet era. Otherwise, there is a real danger of even the so-called major Indic languages becoming extinct soon.

Currently, the recognition engine for Tamil is fairly mature and hence, the next time I write for a Tamil magazine, I am sure I shall be able to use my Tamil recognizer for the same.

A. G. Ramakrishnan

Department of Electrical Engineering,
Indian Institute of Science,
Bangalore 560 012, India
e-mail: ramkiag@ee.iisc.ernet.in