Standardize fonts to catapult Tamil to the digital forefront

A G Ramakrishnan

Department of Electrical Engineering, Indian Institute of Science, Bangalore, India.

Abstract

A strong case is made for the standardization of a few Tamil fonts both for official publications (including school and college books) of the Governments and public display boards (including hoardings, road signs, name boards of establishments) in countries, where Tamil is (one of) the official languages of the State. The basic purpose is to guarantee a great recognition performance for both optical character recognizers and camera based document analysis and recognition systems. In this internet and mobile era, this committed action is bound to tremendously increase access to information for all through the search media, as well as boost the accessibility to all kinds of print information to people with visual disability. With the use of the ubiquitous smart phones, this also can give immediate access to translation of any Tamil material (printed or camera captured) to all the valuable tourist and business visitors. If we are the first in the world to do this, then we can catapult Tamil language to the digital forefront.

Preamble

Standardization is the sine qua non of twenty first century world living, which has made affordable and ultimate technology available to common man. Unless one observes deeply, it may not be immediately obvious to the common man that the many tools and appliances that have become part of our daily lives are attributable to the innumerable standards that have been created by extended discussions and common agreement between the industrial houses, scientists and Governments, as the case may be. For example, in spite of the fact that some of us may have issues with the current Unicode standard, no one can deny the tremendous increase it has provided to access of information on the internet. This is in general true, that even an imperfect standard is a lot better than near perfect, multiple technologies that do not communicate to each other. This is also immediately evident today, where the key inputs, representation, transmission and rendering of Tamil text has not been standardized across the many manufacturers of mobile phones of the world. Partly, this is also due to the lack of adequate pressure on the involved parties by the concerned public and Government machinery. Because of this, we are not able to do the simple thing of texting in Tamil through our mobiles in an era, where one can take a colour photo and send it anywhere in seconds!

Gains due to standardization of communication technology

Without standardization of the modulation/demodulation, transmission technology, as well as the design and manufacturing technology of VLSI processors, we won’t be enjoying the amazing benefits of talking to anyone in the world by the click of a button. Standardization of the several layers of the internet protocol has connected the world to an unbelievable level, that experts in any field (say medicine or art or archaeology) are able to instantly share their new knowledge, as well as benefit from the knowledge of other experts across the globe, by posting their questions on the internet discussion forums. Standardization of medical image formats through DICOM has resulted in considerable progress in the area of medical image processing, which has enabled sharing of crucial diagnostic information between different imaging modalities. This has increased possibilities for information fusion from the multiple modalities such as x-ray computed tomography (CT) and magnetic resonance imaging (MRI), thus facilitating improved diagnosis to the benefit of both patients and physicians.

Negative Examples

We will now see some examples as to how non-standardization leads to difficulties for the people. In India, there are several educational boards that have distinct syllabi and method of evaluation at the school level: State Boards, Cental Board of Secondary Education (CBSE) and Indian Council of Secondary Education (ISCE). Because of these differences, when these students apply for admission to colleges offering engineering education, they are asked to write separate entrance examinations, based on the performance in which these candidates are given admission in the college. Once again, there is difference of opinion between different deemed Universities and the State Governments with respect to these entrance examinations, because of which one needs to write multiple entrance examinations!

Virtual Education

Standardization of Unicode has made every Tamil website and digital document accessible to everyone and information exchange and dissemination has been highly facilitated. Standardization of image formats through jpeg, tif, etc. has resulted in easy sharing of photographs and other pictures through internet. Standardization of video formats has resulted in the revolutionary phenomenon of youtube, which has facilitated rapid sharing of very valuable information such as educational videos. Many reputed Universities of the world have made available open courseware through the internet, which has extended the reach of great teachers to committed students anywhere in the world. The availability of audiovisual pedagogy on demand, as opposed to one’s reading of text from a book, has enhanced the possibility of rapid learning.

e-Shopping

In today’s automated world, we have standardized the dimensions of our dresses too, which has made easy and affordable availability of readymade outfit. In fact, one is able to buy clothes, shoes and other bodyware online today, only because of standardized dimensions, image formats, secure online payments and international standardization of plastic money. The introduction of standardized sizes of doors have made replacement of damaged doors easy. In the same way, standardization of fonts for Government documents, forms and school books will go a very long way in enabling easy access to information through optimized OCR technology for the people with visual disability and even the majority population.

Enhanced interaction between human beings

In the internet era, use of mobiles has become ubiquitous for various applications such as sending emails, finding routes, seeing the meaning for a word, listening to music, making railway and bus reservations, taking photographs and short videos. Standardization of one or two good Tamil fonts for road displays, mile stones, billboard and signboards is bound to rapidly lead to reliable technology for mobile based capture, recognition and transliteration or translation of text. While we can happily use Tamil for all personal and official communications, it will not be a barrier for our friends, business associates and visitors, who read and speak other languages of the world.

Effectiveness of postal automation through standardization

In the United States of America, the sizes of the postal envelopes have been standardized to automate sorting efficiently through machines. Non-standard sizes of postal articles are penalized by an enhanced rate of postage.

Camera based document analysis and recognition

In today’s world, there is too much of interdependence between the states and even countries in terms of economy, business, pleasure travel, etc. Outsourcing of various services has also necessitated wide spread travel of personnel to new places, where they may not know the local language and/or script. A unique phenomenon in India is that the language spoken changes every time and in every direction one travels around 500 km. This is not true even when one travels 1000 km in USA, Russia or China. Accessibility to information for business and tourist travellers and hence workability will be enhanced considerably, if the traveller can read and/or translate information in the foreign language (in our context, Tamil) to her/his language. Similarly, when a Tamil native is in other territories, she/he can use this technology to quickly translate necessary information in the local language to Tamil. Technology exists today, that can detect, localize [1] and extract [2] text reliably from images captured on the road using digital or mobile camera (scene images). Even if this text is curved or wavy, we can detect the contour and align the text line image to be horizontal [3], before attempting to recognize the text. Currently, since the font size, style and background are not standardized, the algorithm needs to have special capabilities [4] to deal with such images. This has necessitated active research in this area [5] and creation of many standard annotated databases [6] and tools [7]. Many of these will be obviated if we standardize the fonts, font sizes, and also foreground and background colors of the road signs, display boards, etc.

For example, in restaurants, one can (i) capture the menu items on the mobile camera, recognize it and quickly translate it to one's own language; (ii) identify names of shops and establishments; (iii) effortlessly identify the destinations of buses and trains; (iv) follow signs in airports and so on.

Recognition of printed text

For the past decade, we have been engaged in developing technology to digitize school and college books [8,9] for the benefit of students with visual disability. Using our OCR technology, close to 200 Tamil school, college and other books have been converted to Braille books by Worth Trust, Chennai, which are being regularly used by blind students. Indian law has made amendments to the copyright law, so that any book can be converted for the exclusive use of blind people, without violating the copyright laws. However, when different books use different fonts and formats, one is not able to always assure the best possible recognition performance by an OCR. However, at least for the school and college books published or authorized by the Government machinery, if the page layout and fonts are standardized, then the OCR can be optimized to give ultimate performance for such standard books. This will facilitate rapid digitization of existing books for such purposes. Similarly, if the court proceedings, real estate registration documents, Government notifications, forms, etc. are in standardized Tamil fonts, then any time the need arises, they can be digitized with great accuracy and speed.

Optical character recognizers can give a very high performance, if they are tuned to one or more specific fonts. While one may prefer to use artistic fonts for certain personal documents such as marriage invitations and award citations, it is highly desirable that we standardize fonts for all displays meant for the public, such as road names, bus and train boards. One can quickly capture such information using a mobile camera, recognize the text and transliterate or translate it to the target language. Since many of these boards contain common and proper nouns, even transliteration will go a long way in giving the essential information needed by a visitor. However, since the text involved is usually small, even effective translation may not present a great technical challenge.

Conclusion

If the merits available are sufficiently understood, then through committed planning of the Governments, the key players can be educated and the entire arena transformed in no time. Let us together cause this happen to Tamil first in the world; then we can catapult Tamil language to the digital forefront in the Universe.

References

1. T. Kasar and A. G. Ramakrishnan, “Multiscript and multioriented text localization from scene images,” Proc. 4th International Workshop on Camera-based Document Analysis and Recognition (CBDAR 2011), pp. 1-14, 2011, Beijing, China.
2. T. Kasar and A. G. Ramakrishnan, “COCOCLUST: Contour-based color clustering for robust text segmentation and binarization,” Proc. 3rd workshop on Camera-based Document Analysis and Recognition (CBDAR 2009) , pp. 11-17, 2009, Spain.
3. T. Kasar and A. G. Ramakrishnan, “A method to extract and align text of an image captured and a system thereof,” Indian Patent Office Reference. No: 109/CHE/2011, 2011.
4. Deepak Kumar, M. N. Anil Prasad and A. G. Ramakrishnan, “NESP: Nonlinear enhancement and selection of plane for optimal segmentation and recognition of scenic word images,” Proc. International Conference on Document Recognition and Retrieval(DRR) XX, 5-7 February 2013, San Francisco, CA USA.
5. Deepak Kumar and A. G. Ramakrishnan, “Power-law transformation for enhanced recognition of born-digital word images,” Proc. 9th International Conference on Signal Processing and Communications (SPCOM 2012), 22-25 July 2012, Bangalore, India.
6. Deepak Kumar, M. N. Anil Prasad and A. G. Ramakrishnan, “Benchmarking recognition results on camera captured word images datasets,” Proc. Workshop on Document Analysis and Recognition (DAR 2012), 16 December 2012, IIT Bombay, Mumbai, India.
7. T. Kasar, D. Kumar, M. N. Anil Prasad, D. Girish, A. G. Ramakrishnan, “MAST: Multi-script annotation toolkit for scenic text,” Proc. Joint Workshop on Multilingual OCR and Analytics for Noisy and Unstructured Text Data (J-MOCR-AND 2011), pp. 1-8, 2011, Beijing, China.
8. Peeta Basa Pati and A. G. Ramakrishnan, “Word level multi-script identification,” Pattern Recognition Letters, 2008, Vol. 29, pp. 1218-1229.
9. D. Dhanya and A. G. Ramakrishnan, “Optimal feature extraction for Bilingual OCR,” Proc. Fifth IAPR Workshop on Document Analysis Systems DAS-02, Princeton, NJ, August 19-21, 2002, pp.25-36.