A Database for Handwriting Recognition Research in Sinhala Language

H. C. Fernando[‡] chandrika@sliit.lk
N.D. Kodikara[†] nihal.uoc@mail.cmb.ac.lk
S. Hewavitharana[†] sanjika@cmb.ac.lk

[‡]Sri Lanka Institute of Information Technology, Colombo, Sri Lanka
[†] University of Colombo School of Computing, Colombo, Sri Lanka

Abstract

This article presents a database of images of handwritten city names. The aim is to provide a standard database for Sinhala handwriting recognition research. This database contains about 15,000 images of about 500 city names of Sri Lanka. These images are obtained from the addresses of live mail so that the writers had no idea that they would be used for this purpose. Also, these are unconstrained handwriting images unlike the images collected using prescribed forms in laboratory environment. The images are divided into two groups, training set and testing set. This enables the comparison of results of different researches and serves the purpose of being a standard database.

1. Introduction

Character and handwriting recognition has a great potential in data and word processing, for instance, automated postal address and ZIP code reading, data acquisition in banks, text-voice conversions etc. As a result of intensive research and development efforts, systems are available for English language [1], [2], [3], [4], Chinese language [5], Japanese language[6], and handwritten numerals [7].

There is still a significant performance gap between the human and the machine in recognizing unconstrained handwriting. This is a difficult research problem caused by huge variation in writing styles and the overlapping and the intersection of neighboring characters of varying degree of neatness from very sloppy to neat [8].

Early researches in Sinhala handwriting recognition have used datasets specifically collected for the particular research. Most of them have used constrained handwriting forcing the writer to write on a ruled paper [9],[10]. Due to the varying nature of the datasets used for training, the recognition results can hardly be compared. Therefore, a standard database of images is needed for research in handwriting recognition. This database must contain all the information so that the researcher can employ different pre-processing systems. For example, most of such databases for other scripts contain only binary images. Therefore, experimentation with pre-processing is not possible. The information is limited by the thresholding and the sampling rate provided by the digitizer used. Recognition accuracy can be improved by using gray-level images since they contain more information. Another important feature is to have unbiased data. If the subjects are aware that the handwriting is collected to develop automatic recognition algorithms then their handwriting could be biased. They may write with abnormal neatness.

A standard database should also contain precisely defined training and testing sets. This would facilitate comparison of results that are caused by minor differences in testing data themselves. This will also help to minimize replication of experiments and the comparison of performance. Furthermore, a common database will allow the researchers to share test results. Algorithms of different researchers can be combined to obtain better results faster.

This article describes a database that could be used as a standard database for handwritten text recognition research of Sinhala language. This contains about 15,000 Sinhala words. They are city names of about 500 cities in Sri Lanka obtained from the addresses of live mail. There are two types of cities in Sri Lanka when delivery of mail is considered, namely, cities with Post Offices that handle delivery of mail and cities with Sub Post Offices that do not handle delivery of mail. There are about 500 cities that belong to the first category, that is, Post Offices that handle delivery of mail. All of them were included with the intention of using them in future for a possible automation of mail sorting. Thirty images of each city were obtained to compile the database of 15,000 images. The value 30 was chosen for maintaining Normal Distribution in features extracted for classification. Live mail was considered to be the most suitable to obtain unconstrained handwriting because the writers had no

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

Table 1. Distribution of characters in the database

Training Dataset								
අ	ආ	ඇ	ඉ	ඊ	උ	ඌ	එ	ඒ
735	53	210	210	24	370	105	50	24
ක	ඛ	ඕ	ඛ	ග	ච	ජ	ඤ	ඩ
2760	210	105	3620	4600	370	131	50	1995
ද	ධ	ණ	ප	බ	ම	ව	ළ	න
2870	105	525	3130	1320	4540	1630	6485	4515
ය	ර	ෂ	ව	ස	හ	ං	ඔ	
2730	4170	230	4805	1285	2070	315	470	
Testing Dataset								
අ	ආ	ඇ	ඉ	ඊ	උ	ඌ	එ	ඒ
105	7	30	30	6	50	15	10	6
ක	ඛ	ඕ	ඛ	ග	ච	ජ	ඤ	ඩ
390	30	15	520	650	50	19	10	285
ද	ධ	ණ	ප	බ	ම	ව	ළ	න
400	15	75	440	180	650	230	925	645
ය	ර	ෂ	ව	ස	හ	ං	ඔ	
390	600	40	685	185	300	45	70	

idea that the addresses would be used for this purpose. Mail address also gives the required degree of variation into the dataset and does not impose any restriction to the writer when writing the characters. Sampling frame was the list of post offices given in the book of postal codes published by the Sri Lanka Postal Department. The city names were sampled from this list. We also tried to cover as many letters as possible of the alphabet. The database is also divided into two sets of data, namely, training set and testing set.

2. Sinhala Language

Sinhala is the language used by the majority of the people in Sri Lanka. It is a member of the Indo-Aryan family of languages along with Sanskrit, Hindi and Bengali. Sinhala script, which is alphabetic in nature, is derived from ancient North Indian script Brahmi and it is unique to this language.

The contemporary Sinhala alphabet contains 18 vowels, 2 semi-vowels and 41 consonants. The vowels are inherent in consonants, so the number of different symbols for vowel-consonant compositions is quite high. However, the database considered here contains 11 out of 18 vowels (60%) and 22 out of 41 consonants (54%), as the others are not found in city names.

3. The Database

The database can be searched for any Sinhala character given in Table 1. The list of all the city names that contain the character of interest will appear on the screen. Then, any of the cities in the list can be chosen. The output would be the 30 images collected for that particular city.

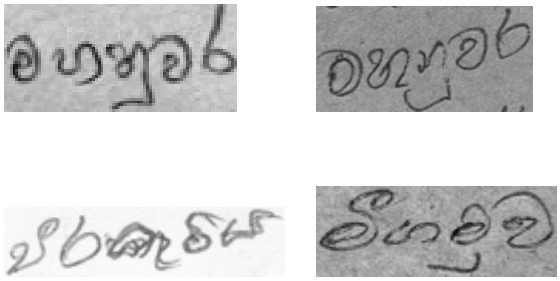

Figure 1. Some city name images

envelopes from Sri Lanka Central Mail Exchange and represented in the database with an overall view of developing an algorithm to recognize Sinhala handwritten characters. The database is classified according to the first letter of the city name. Some of the city name images are given in the Figure 1.

4. Summary and Conclusions

The database presented here has no subject-bias problems like other databases that were prepared in laboratories. The images are 8-bit gray scale and the researchers can try various methods of pre-processing and feature extraction. When the images are binarized, it leads to loss of information and leads to unsatisfactory feature extraction. This in turn results in low rates of recognition.

This database provides handwriting in a restricted domain, namely city names. But, this is a realistic simulation for handwritten research in a wider spectrum of applications.

The database is available with the funding organization, the National Science Foundation (NSF) of Sri Lanka, so that the researchers can access the database.

Appendix: Database Specifications

Medium: The data is provided on an ISO-9660 format CDROM. The CDROM is readable on either a PC or a workstation equipped with the appropriate hardware and driver software.

Format: The images are represented in bitmap format with 300 pixels/inch in 8-bit grey-scale.

Image Digitizer: All the images were scanned on a HP Scanjet 5200C digitizer.

The Interface Software: The database is available with a user-friendly interface so that the users can access the character images according to alphabetical order of the city names.

Acknowledgement

National Science Foundation of Sri Lanka funded this project (Grant No. RG-2001-IS-02). This project was inspired by a similar project conducted by the Centre of Excellence for Document Analysis and Recognition (CEDAR) at the University of New York at Buffalo for automation of postal service in the United States of America.

References

[1] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, pp. 68-83, Jan. 1989.

[2] J. Hu, M. K. Brown and W. Turin, "HMM based on-line handwriting recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, pp. 1039-1045, Oct. 1996.

[3] G. Kim and V. Govindaraju, "A lexicon driven approach to handwritten word recognition for real-time applications," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, pp. 366-379, April 1997.

[4] R. Buse, Z-Q Liu, and T. Caelli, "A structural and relational approach to handwritten word recognition," IEEE Trans. Systems, Man and Cybernetics, Part-B, vol. 27, pp. 847-861, Oct. 1997.

[5] K. Liu, Y. S. Huang and C. Y. Suen, "Identification of fork points on the skeletons of handwritten Chinese characters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, pp. 1095-1100, Oct. 1999.

[6] M. Okamoto and K. Yamamoto, "On-line handwriting character recognition using direction change features that consider imaginary strokes," Pattern Recognition, 32, pp. 1115-1128, 1999.

[7] J. Cai and Z-Q Liu, "Integration of structural and statistical information for unconstrained handwritten numeral recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, pp. 263-270, March 1999.

[8] J. J. Hull, T. K. Ho, J. Favata, V. Govindaraju, and S. N. Srihari, "Combination of segmentation-based and holistic handwritten word recognition algorithms," Proc. From Pixels to Features III, Int. Workshop Frontiers Handwriting Recogn., Bonas, France, pp. 229-240, 1991.

[9] S. Hewavitharana and N.D.Kodikara, "A Statistical Approach to Sinhala Handwriting Recognition", Proc. of the International Information Technology Conference (IITC), Colombo, Sri Lanka, Oct. 2002.

[10] S. Hewavitharana, H. C. Fernando and N.D. Kodikara, "Off-line Sinhala Handwriting Recognition using Hidden Markov Models", Proc. of Indian Conference on Computer Vision , Graphics & Image Processing (ICVGIP) 2002, Ahmedabad, India, pp. 266-269, Dec. 2002.