Advances in Information Mining, ISSN: 0975–3265, Volume 2, Issue 2, 2010, pp-11-13

Handwritten Marathi character (vowel) recognition

Ajmire P.E.[1] and Warkhede S.E.[2]
[1]G. S. Sci., Arts & Comm. College, Khamgaon, peajmire@rediffmail.com
[2]S.S.A, Z.P.Buldhana, ssabuldhana1@yahoo.co.in

Abstract- The work in this paper is deals with the recognition of handwritten Marathi vowels. Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighbouring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtri, one of the Prakrit languages which developed from Sanskrit. This work is based on invariant moments for recognition of isolated Marathi Handwritten Characters and their divisions. The proposed technique is independent of size, slant, orientation, translation and other variations in handwritten characters. Handwritten Marathi Characters/Numerals are more complex for recognition than corresponding English characters due to many possible variations in order, number, direction and shape of the constituent strokes. The work treats isolated Characters as an image of 40X40 pixel size. Seven invariant central moments of the image and two additional feature sets are evaluated. 10 samples of each number from 20 different writers have been sampled and prepared a database has been made. Seven central invariant moments are evaluated for each character and its parts by dividing it by two different ways. In all, there are 14 features corresponding to each character. The Gaussian Distribution Function has been adopted for classification. The average success rate of some vowels is compatible.
Keyword: Handwritten character recognition, invariant moments.

Introduction

Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighbouring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtri, one of the Prakrit languages which developed from Sanskrit. Marathi first appeared in writing during the 11th century in the form of inscriptions on stones and copper plates. From the 13th century until the mid 20th century, it was written with the Modi alphabet. Since 1950 it has been written with the Devnagari alphabet. Recognition of handwritten characters has been popular research areas since 1870. There are many script and Languages in the world. The researchers have done work on some of them like English, Chinese, Latin, Arabic [1], Japanese [2], Thai [3] and Devnagari [4].

There are many scripts and languages in India and not much work has been done for recognition of handwritten Marathi character. Recognition of handwritten characters is important because of its applicability to a number of problems, like postal code recognition and information extraction from fields of different forms. In the Indian context, there exists a need for development and/or evaluation of the existing techniques for recognition of numerals written in Indian scripts. Generic techniques cannot, in general, tackle problems associated with script specific peculiarities. In this paper we propose a system towards the recognition of off-line handwritten Marathi numeral. There are many pieces of work have been done towards the recognition of printed Devnagari Character[4] and at present OCR systems are commercially available for some of the printed Indian scripts, many research has been carried out for Bangala[5]. This work is based on invariant moments for recognition of isolated Marathi Handwritten Characters (Vowels). The proposed work is combination of language and Computer Science. This is the application of Computer in languages (scripts). Thus it is towards the development of Marathi language using applied science and emerging technologies. The subject also brushes to statistics and mathematics. The significance of the studies is in language, administration for e-governance, etc. Thus the problem it self is multidisciplinary.

Marathi was written in Modi script — a cursive script designed for minimising the lifting of pen from paper while writing. Most writings of the Maratha Empire are in Modi script. However, Persian-based scripts were also used for court documentation. With the advent of large-scale printing, Modi script fell into disuse, as it proved very difficult for type-setting. Currently, due to the availability of Modi fonts and the enthusiasm of the younger speakers, the script is far from disappearing.

Now, Marathi is written in the Devnagari script, an alphasyllabary consisting of 16 vowels and 36 consonants making a total of 52 letters. It is written from left to right. Devnagari used to write Marathi is slightly different than that of Hindi or other languages. Earlier, another script called 'Modi' was in use till the time of the Peshwas(18th century). This script was introduced by Hemad Panta, a minister in the court of the Yadava kings of Devgiri (13th century). This script looked more like today's draviDian scripts and offered the advantage of greater writing speed because the letters could be joined together. Today only the Devnagari script is used which is easier to read but does not

have the advantage of faster writing. The script currently used in Marathi is called 'Balbodh' which is a modified version of Devnagari script. From this script, Marathi language has 12 common vowels and for this the database is created for these vowels. Vowels are the soul of the speech and sounds of the language. In Marathi vowels are usually written in abbreviated symbols or forms.

Design Goal

Automatic handwriting recognition has been studied for a long time, because of its widespread use as a means for human-to-human communication. Its generation and recognition aspects served as a paradigm in order to study models and generic techniques, besides the need to solve the specific problem. On the other hand, its variability, due to differences in cultures, writer styles, alphabets or fonts, makes it an interesting problem in the field of pattern recognition[5]. Its two-way relationship with the humans (generation and perception) requires an involvement of knowledge and techniques, more related to psychology. Advancements in electronics allowed the computing power necessary for handwriting recognition to fit into a smaller form factor than tablet computers, and handwriting recognition is often used as an input method for hand-held PDAs. The first PDA to provide written input was the Apple Newton, which exposed the public to the advantage of a streamlined user interface. However, the device was not a commercial success, owing to the unreliability of the software, which tried to learn a user's writing patterns. Despite many years of research in the field of handwriting recognition technology, IT has not reached the masses in many of the local languages. Marathi is the one for instance. One can imagine how much easier the people's lives will be if a portable machine can understand what one writes in natural handwriting mode. It is certainly difficult to type addresses, messages, memos and important information etc. by using existing computer for those who are non-native to English, computer novices or feel inconvenience in using a keyboard and keypad(old people). In such a case, writing would be more clear and easy to understand in their own local languages. Therefore, a system having the intelligence in recognizing the natural handwriting is a desperate demand in the global market. Marathi keyboards are cumbersome to use. However, new pen tablets offer the possibility of online handwriting, when combined with handwriting recognition technology. Indian Language Technology Vision 2010 has been prepared with the Vision statement "*Digital Unite and Knowledge for All*". Technology audit focuses on peer-review, peer-technology sharing and product–oriented technology development.

Database design Description

For this work, the handwritten database is prepared by scanning the images of handwritten vowels. These vowels are written by different people. The sample images of some vowels are given in figure 1.

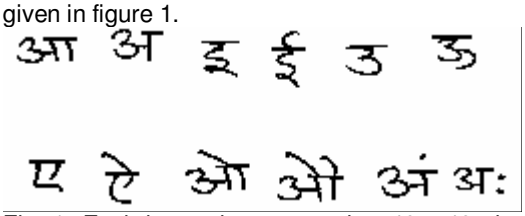

Fig. 1- Each image is converted to 40 x 40 pixel size. The images converted to Black and White by using MATLAB. The database of these images was used for this work of recognition.

Feature Extraction.

The purpose of feature extraction is to reduce image data by measuring certain features or properties of each vowel. R. J. Ramteke, P. D. Borkar, S. C. Mehrotra [6] describes an approach based on invariant moments for recognition of isolated Marathi Handwritten Numerals and their divisions. The proposed technique is independent of size, slant, orientation, translation and other variations in handwritten characters. Handwritten Marathi Characters/Numerals are more complex for recognition than corresponding English characters due to many possible variations in order, number, direction and shape of the constituent strokes. The work treats isolated vowels as an image of 40X40 pixel size. Seven invariant central moments of the image and two additional feature sets are evaluated. 10 samples of each vowel from different writers have been sampled and prepared a database. Seven central invariant moments are evaluated for each character. We use these features drawn by invariant moment technique. The invariant moments are well known to be invariant under translation, scaling, rotation and reflection [7,8]. They are measures of the pixel distribution around the centre of gravity of the character and allow to capture the global character shape information. Traditionally, moment invariants are computed based on the information provided by both the shape boundary and its interior region. The moments used to construct the moment invariants are defined in the continuous but for practical implementation they are computed in the discrete form. Given a function *f(x,y)*, these regular moments are defined by:

$$M_{pq} = \iint x^p y^q f(x,y)\,dx\,dy$$

Where, M_{pq} is the two-dimensional moment of the function *f(x y)* for *p*, *q* = 0,1, 2, …. The order of the moment is *(p+q)* where *p* and *q* are both

natural numbers. For implementation in digital form this becomes:

$$M_{pq} = \sum_{X} \sum_{Y} x^p y^q f(x, y)$$

To normalise for translation in the image plane, the image centroids are used to define the central moments. The co-ordinates of the centre of gravity of the image are calculated using equation and are given by.

$$\mu_{pq} = \sum_{X} \sum_{Y} (x - \bar{x})^p (y - \bar{y})^q$$

A set of seven moment invariant values ($\phi\tilde{1}$ - ϕ_7) is derived by equations in[9,10] which are used as features. The moment invariants are evaluated by taking absolute value of log of the moment of each of the sample image. Further, mean and standard deviation are determined for each feature using 120 samples. Thus we had 14 features (7 means and 7 standard deviations), which are applied as features for recognition using Gaussian Distribution Function.

Result and Conclusion

The system was tested on 12 handwritten Marathi vowels and 10 images for each vowel.

Table 1- Following table shows the recognition. To improve the recognition rate we use Gaussian distribution function for clustering of the vowels as given below,

Cluster 1 content	A	AA	O	AU	AM	Aha
Cluster 2 content	I	EE				
Cluster 3 content	U	OO				
Cluster 4 content	AE	AI				

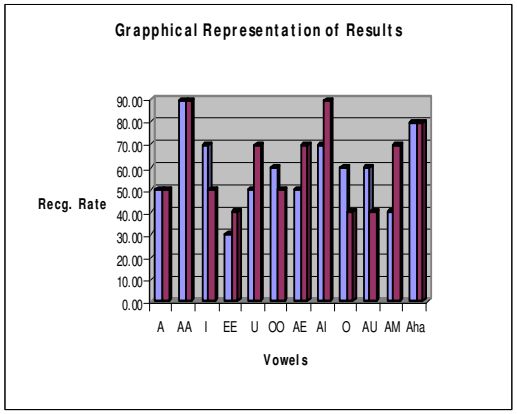

Fig. 2- The over all average recognition rate of vowels before clustering is 59.% and after the clustering is 62%. The graphical representation before and after clustering is given below.

Hence we can say that the clustering improves the recognition rate.

References

[1] Jannoud I.A. (2007) *American Journal of Applied Sciences* 4 (11): 857-864.

[2] Tour Wakahara, Y. Kimura & Mutsuo (2001) *Handwritten Japanees Character Recognition Using Adaptive Normalization by Global Affine Transformation. Proc. 6th ICDAR Vol., Issue*.

[3] Mitrpanont J.L., Limkonglap U. (2007) *Using Countour Analysis to improve Feature Extraction in Thai Handwritten Character Recognition Systems. Proc. 7th IEEE ICCIT*.

[4] Pal U., Sharma N., Wakabayashi T. and Kimura F. (2007) *Off-line Handwritten character recognition of Devnagari Script. 9th ICDAR*.

[5] Bhowmik T.K., Roy A. and Roy U. (2005) *Character Segmentation for Handwritten Bangla Words using Artificial Neural Network. Proc.1st IAPR TC3 NNLDAR*.

[6] Ramteke R.J., Borkar P.D., Mehrotra S.C.(2005) *Proc. of International Conference on Cognition and Recognition (ICCR 2005), Mysore, (Karnataka), India*, 482 - 489.

[7] Swethalakshmi H., Anitha Jayaraman, Srinivasa Chakravarthy V., Chandra Sekhar C. (2006) *Tenth International Workshop on Frontiers in Handwriting Recognition*, HAL: inria-00104402, version 1.

[8] Deshpande P.S., Latesh Malik, Sandhya Arora (2007) *Handwritten Devnagari Character Recognition Using Connected Segments And Minimum Edit Distance. IEEE 2007*.

[9] Ajmire P.E., Talwankar D.S. and Ramteke R.J. (2008) *Recognition of Fossil Leaves Genera (Lower Gondwana Flora). Proc.of 1st Int. Conf. on ACVIT,* 8, 1302-1306.

[10] Ramteke R.J., Mehrotra S.C. (2006) *Feature Extraction Based on Invariants Moment for Handwriting Recognition, Proc. of 2nd IEEE Int. Conf. on Cybernetics Intelligent System (CIS2006), Bangkok*.