International Journal Of

Recent Scientific Research

ISSN: 0976-3031

Volume: 7(1) January -2016

PERFORMANCE AND ANALYSIS OF HANDWRITTEN TAMIL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORKS

Rajasekar M., Celine Kavida A and Anto Bennet M

THE OFFICIAL PUBLICATION OF
INTERNATIONAL JOURNAL OF RECENT SCIENTIFIC RESEARCH (IJRSR)
http://www.recentscientific.com/ recentscientific@gmail.com

Available Online at http://www.recentscientific.com

ISSN: 0976-3031

International Journal of Recent Scientific Research
Vol. 7, Issue, 1, pp. 8611-8615, January, 2016

International Journal of Recent Scientific Research

RESEARCH ARTICLE

PERFORMANCE AND ANALYSIS OF HANDWRITTEN TAMIL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORKS

Rajasekar M[1]., Celine Kavida A[2] and Anto Bennet M[3]

[1]Department of CSE, Veltech Multitech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai-600062, India

[2]Department of Physics, Veltech Multitech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai-600062, India

[3]Department of ECE, VEL TECH, Chennai-600062, India

ARTICLE INFO

Article History:

Received 16th October, 2015
Received in revised form 24th November, 2015
Accepted 23rd December, 2015
Published online 28th January, 2016

Key words:

Optical Character Recognition (OCR), Computer handwritten character recognition (HCR), Artificial Neural Networks (ANN),

ABSTRACT

Computer handwritten character recognition (HCR) system can improve the human computer interaction and better integrate computers into human society. HCR and optical character recognition (OCR) in a more general context are an integral part of pattern recognition. At the early stages of research and development of pattern recognition most of the researcher investigated the subject of OCR. One of the main reason was because characters were very handy to deal with, since most of the time characters are defined in a two dimensional lattice which have two states, so it was commonly thought that this problem could be easily solved. However, against what was the expectation after some initial progress, great difficulty in solving this problem surfaced. And even today with large scale computer power available and high-quality scanners, OCR still poses some interesting and difficult problem to be solve definitely. Another point of interest is that optical character recognition is rather a universal problem in that it includes essential problems of pattern recognition which are common to all other topics. Thus making it an interesting problem to analyze in views of understanding other more complex problems in pattern recognition and analysis. In this proposed work, introduce the problem of handwritten Tamil character recognition including a historical background on the subject. A review of computer handwritten recognition aims and application is studied, followed by description of previous method and techniques. Handwritten Tamil character recognition, two different approaches in trying to deal with this problem is studied. A moment based feature extraction technique and a coding scheme based on the neighborhood relation are developed. The fuzzy c-means clustering algorithm is used as the method for data reduction. And Artificial Neural Networks (ANN) are applied for the recognition process.

INTRODUCTION

Optical Character Recognition (OCR) deals with machine recognition of characters present in an input image obtained using scanning operation. It refers to the process by which scanned images are electronically processed and converted to an editable text. The need for OCR arises in the context of digitizing Tamil documents from the ancient and old era to the latest, which helps in sharing the data through the Internet. The main aim of our paper is to develop an OCR for alphanumeric and Tamil text. This is implemented as a MATLAB program using image processing toolbox. MATLAB programming contains three modules namely preprocessing, segmentation and recognition. Tamil language has 12 vowels and 18 consonants. These are combined with each other to yield 216 composite characters and 1 special character (aayutha ezhuthu) counting to a total of (12+18+216+1) 247 characters.

English has 5 vowels and 21 consonants characters counting to a total of 26 characters. Numbers from 0 to 9 count to a total of 10 different characters. Optical Character Recognition (OCR) refers to the process of converting printed documents into software translated Text. These documents available in the form of books, papers, magazines, etc. are scanned using standard scanners which produce an image of the scanned document. Now-a-days OCR'S are available for documents in English Language using various character recognition methods. It is available to the end user for transforming printed English

Corresponding author:* **Rajasekar M
Department of CSE, Veltech Multitech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai-600062, India

books to editable e-books. But OCR's for Tamil documents are not available as a whole.

LITERATURE REVIEW

There are OCR'S available for languages like Arabic, kannadam, Chinese,etc. Most of these OCR'S including that of English requires a Neural Network for their implementation. Training a Neural Network for implementing an OCR is time consuming. So we may go for alternatives like Unicode mapping, correlation methods for character recognition.As far as Unicode mapping is concerned we need to extract the features of the characters like height, width, the number of horizontal and vertical lines (long and short) , the horizontally and vertically oriented curves, circles, slope lines, image centroid and special dots. These leads to more number of computations and there by decreases the execution speed. So we go for correlation method in which each character is compared with the characters stored in databases and recognized

Proposed System

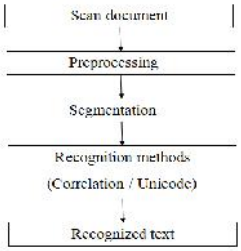

Fig 1 Ocr Functional Block Diagram

Scanning

A properly printed document is chosen for scanning. It is placed over the scanner. A scanner software is invoked which scans the document. The document is sent to a program that saves it in preferably TIF, JPG or GIF format, so that the image of the document can be obtained when needed. The size of the input image is as specified by the user and can be of any length but is inherently restricted by the scope of the vision and by the scanner software length shown in fig 1.

Preprocessing

This is the first step in the processing of scanned image. The scanned image is checked for skewing. There are possibilities of image getting skewed with either left or right orientation. Here the image is first filtered and then binarized. We use median filter for noise elimination. The function for skew detection checks for an angle of orientation between ±15 degrees and if detected then a simple image rotation is carried out till the lines match with the true horizontal axis, which produces a skew corrected image. Skew correction is done by rotating the image around an angle.

Segmentation

After pre-processing, the noise free image is passed to the segmentation phase, where the image is decomposed into individual characters.

Algorithm for segmentation

- The binarized image is checked for inter character spacing.
- If inter character spaces are detected then the image is segmented into sets of characters.
- Then each characters and their common background are labeled for easing the recognition phase.

Recognition methods

Correlation

Correlation indicates the strength and direction of a linear relationship between two random variables. That is in contrast with the usage of the term in colloquial speech, denoting any relationship, not necessarily linear. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data.

Given a set of observations (x_1, y_1), (x_2,y_2),...(x_n,y_n), the formula for computing the correlation coefficient is given by

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \quad \text{---------- (1)}$$

The correlation coefficient always takes a value between -1 and 1, with 1 or -1 indicating perfect correlation (all points would lie along a straight line in this case). A positive correlation indicates a positive association between the variables (increasing values in one variable correspond to increasing values in the other variable), while a negative correlation indicates a negative association between the variables (increasing values is one variable correspond to decreasing values in the other variable). A correlation value close to 0 indicates no association between the variables. For this reason, the correlation coefficient is often more useful than a graphical depiction in determining the strength of the association between two variables.

Correlation Matrices

The correlation matrix of n random variables $X_1 ... X_n$ is the $n \times n$ matrix whose i,j entry is $\text{corr}(X_i, X_j)$. If the measures of correlation used are product-moment coefficients, the correlation matrix is the same as the covariance matrix of the standardized random variables $X_i / \text{SD}(X_i)$ for $i = 1, ..., n$. Consequently it is necessarily a positive-semi definite matrix.The correlation matrix is symmetric because the correlation between X_i and X_j is the same as the correlation between X_j and X_i.

Unicode mapping

The Unicode standard reflects the basic principle which emphasizes that each character code has a width of 16 bits. Unicode text is simple to parse and process, and Unicode characters have well defined semantics. Hence Unicode is

chosen as the encoding scheme for the current work (Unicode, 2000). After classification the characters are recognized and a mapping table is created in which the unicodes for the corresponding characters are mapped. Table 1 shows one such rule based classifier based on Tamil Unicode.

Feature extraction

This follows the segmentation phase of OCR where the individual image glyph is considered and extracted for features. First a character glyph is defined by the following attributes: (1) Height of the character; (2) Width of the character; (3) Numbers of horizontal lines present—short and long; (4) Numbers of vertical lines present—short and long; (5) Numbers of circles present; (6) Numbers of horizontally oriented arcs; (7) Numbers of vertically oriented arcs; (8) Centroid of the image; (9) Position of the various features; (10) Pixels in the various regions. Detection of character height and character width: This is detected by simply scanning the image glyph and finding the boundary of the glyph in the horizontal and vertical directions. Height=30 and Width=20

Horizontal line detection

Here a mask is run over the entire image glyph and thresholded which detects the horizontal line. The mask is

$$\begin{matrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{matrix}$$

RESULT

No_of_Horiz=1.
Algorithm for Horizontal Line Detection:
c[i][j]=img[i−1][j−1]×(−1)+img[i−1][j]×(−1)
+img[i−1][j+1]×(−1)+img[i][j−1]×2
+img[i][j]×2+img[i][j+1]×2
+img[i+1][j−1]×(−1)+img[i+1][j]×(−1)
+img[i+1][j+1]×(−1)
If (c[i][j]>0)
Increment q[i];
If (q[i]≥smallThreshold)
Small horizontal line is detected
Else
If (q[i]≥longThreshold)
Long Horizontal line is detected;

Vertical line detection

Here a mask is applied over the entire range of pixels and thresholded which detects the vertical line. The mask is

$$\begin{matrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{matrix}$$

RESULT

The number of vertical lines is 1.

Algorithm for vertical line detection:
c[i][j]=img[i−1][j−1]×(−1)+img[i−1][j]×2
+img[i−1][j+1]×(−1)+img[i][j−1]×(−1)
+img[i][j]×2+img[i][j+1]×(−1)
+img[i+1][j−1]×(−1)+img[i+1][j]×2
+img[i+1][j+1]×(−1)
If (c[i][j]>0)
Increment q[i];
If (q[i]≥smallThreshold)
Short vertical line is detected
Else
If (q[i]≥longThreshold)

Long vertical line is detected;

Slope lines detection

The masks for the slope lines are as follows:

$$+45: \begin{matrix} -1 & -1 & 2 \\ -1 & 2 & -1 \\ 2 & -1 & -1 \end{matrix} \qquad -45: \begin{matrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{matrix}$$

To detect slope lines these masks are applied over the image and then thresholded suitably.

Detection of circles and arcs

Here a new mask is derived and operated on each and every pixel in the image glyph and thresholded which detects the circle. A 5×5 mask is taken. The arcs are detected using the circle detection algorithm and checked for the semi-circle and diameter. The mask for circle is

$$\begin{matrix} 2 & 2 & 2 & 2 & 2 \\ 2 & -1 & -1 & -1 & 2 \\ 2 & -1 & -1 & -1 & 2 \\ 2 & -1 & -1 & -1 & 2 \\ 2 & 2 & 2 & 2 & 2 \end{matrix}$$

Character recognition

The scanned image is passed through various blocks of functions and finally compared with the recognition details from the mapping table from which corresponding unicodes are accessed (Table 1) and printed using standard Unicode fonts so that the OCR is achieved (Unicode, 2000).

Table 1 Sample mapping

Features extracted	Classified character	Class label	Unicode
No of short horizontal lines—0		1	0x0b85
No of short vertical lines—0		1	
No of long horizontal lines==1		1	
No of vertically oriented arcs==1			
No of circles==1			
No of horizontally oriented arcs==1			
Height=23			
Width=20			
If input==target then the character is			
– one horizontal arc and one vertical arc		2	0x0b86
		2	
		2	
... Next character			

EXPERIMENTATION RESULTS

Input image

In Matlab Image Processing Toolbox, we read a image using imread () function. We have a Graphical User Interface (GUI) supported by Matlab. It is easy to design a GUI using Matlab, because the source code for a particular GUI will be automatically generated after designing it shown in fig 2.

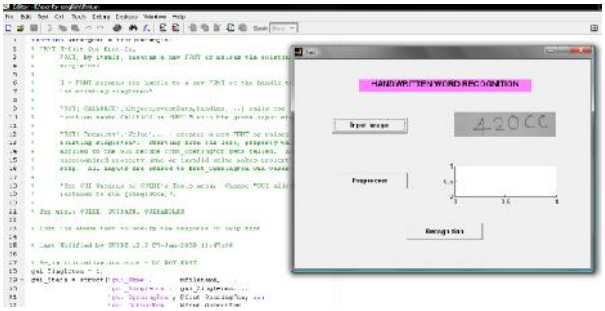

Fig 2 Input image

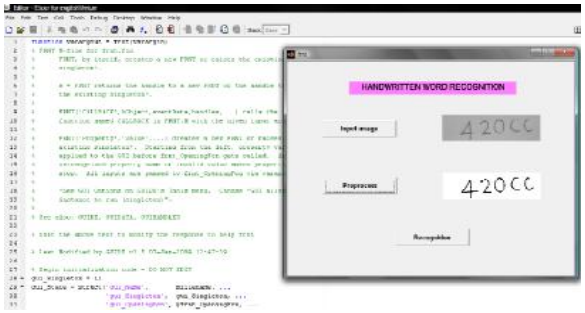

Fig 3 Preprocessing

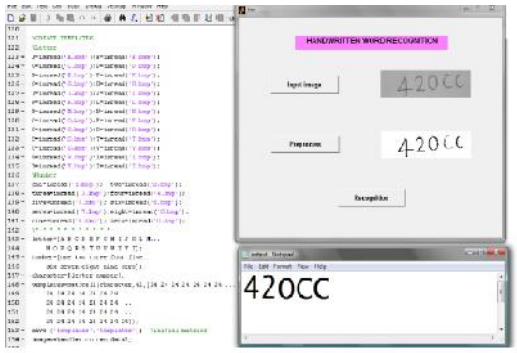

Fig 4 Alphanumeric Text

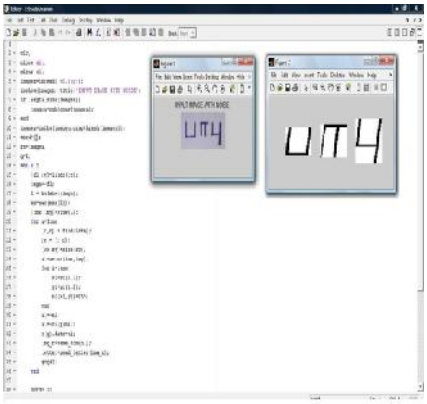

Fig 5 Tamil Text

Preprocessing

The scanned image is preprocessed using edge () function which is used to detect edges by filtering. Here we use Sober filter for improved noise immunity shown in fig 3.

Character Recognition

The preprocessed image of a word is segmented into characters using mat2cell () function (by labeling using font size).We then recognize each characters by correlation methods. OCR is currently used to maintain databases in any organization. If OCR is available then processing and maintaining the student records become easier. The students' forms can be directly scanned, extracted for details and directly transformed into a Student Database. OCR can also play a major role in the business environment. OCR reduces cost and effort by eliminating manual data entry, etc. If OCR is available, it becomes easier to extract and transform the data into business BASE and promote business without the need for large mobility (data, people). The increasing number of faxes and paper documents received by businesses often originate from the same suppliers or customers and has a format and layout that have not changed for some time. The data within these documents have to be manually interpreted and re-keyed into business applications as part of key business processes (e.g. Purchase Orders and Invoices into Accounting Systems for Accounts Receivables and Payables, students data etc.). The larger the volume of documents received, the greater the manual resource required entering the data into business applications. The scope for errors and delay to critical business processes also increases as volume increases, if it is handled manually. By scanning the documents to create TIFF image files and automatically routing electronic fax images to OCR, the errors, cost and delay of manual data entry can be avoided, as OCR can automatically extract data from the documents and format the data for onward delivery to other applications shown in fig 4&5.

CONCLUSION

OCR is aimed at recognizing printed document. The input document is read preprocessed, feature extracted and recognized and the recognized text is displayed in a picture box. Maintaining and getting the contents from and to the books is very difficult. OCR eliminates the difficulty by making the data available in printed format. In a way OCR provides a paperless environment. OCR provides knowledge exchange by easier means. If the knowledgebase of rich contents are created, it can be accessed by people of varying category with ease and comfort. After implementing an OCR for Alphanumeric and Tamil text for few words, an OCR for Alphanumeric and Tamil texts in whole paragraph will be implemented. This will be followed by implementing OCR'S for many languages.

References

1. Dr. AntoBennet, M , Sankaranarayanan S, Sankar Babu G, " Performance & Analysis of Effective Iris

Recognition System Using Independent Component Analysis ", *Journal of Chemical and Pharmaceutical Sciences* 08(03): 571-576, August 2015

2. Anto Bennet, M, Mohan babu, G, Rajasekar, C & Prakash, P, "Performance and Analysis of Hybrid Algorithm for Blocking and Ringing Artifact Reduction", *Journal of Computational and Theoretical nanoscience* vol.12,no.1,pp.141-149,2015
3. AntoBennet, M & JacobRaglend, "Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", *European Journal of Scientific Research*, vol. 66 no.1, pp.120-129, 2011
4. AntoBennet, M & JacobRaglend, "Performance and Analysis of Compression Artifacts Reduction for MPEQ-4 Moving Pictures Using TV Regularization Method", *Life Science Journal* vol. 10, no. 2, pp. 102-110, 2013
5. AntoBennet, M & JacobRaglend, "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", *Journal of Computer Scienc*e, vol. 8, no. 9, pp. 1447-1454, 2012
6. Dr. AntoBennet, M, Sankar Babu G, Natarajan S, "Reverse Room Techniques for Irreversible Data Hiding", *Journal of Chemical and Pharmaceutical Sciences* 08(03): 469-475, September 2015.

****

How to cite this article:

Rajasekar M., Celine Kavida A and Anto Bennet M. 2016, Performance And Analysis of Handwritten Tamil Character Recognition Using Artificial Neural Networks. *Int J Recent Sci Res.* 7(1), pp. 8611-8615.

TSSN 0976-3031