ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

A Bilingual Numeral OCR System for Creating Uni-Lingual Digitized Numeral Document

K. Karthick[1] & S. Chitra[2]

[1] Panimalar Engineering College, Chennai, Tamilnadu, India

[2] Er.Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India

Correspondence: K. Karthick, Assistant Professor (G1) / EEE, Panimalar Engineering College, Nazarathpet, Poonamallee, Bangalore Trunk Road, Chennai, 600123, Tamilnadu, India. Tel: 91-989-482-1761. E-mail: kkarthiks@gmail.com

Received: August 8, 2015 Accepted: August 29, 2015 Online Published: November 30, 2015

doi:10.5539/mas.v9n13p148 URL: http://dx.doi.org/10.5539/mas.v9n13p148

Abstract

The optical character recognition has been used in many applications such as dictionary generation, customer billing system, banking and postal automation, and library automation etc. The bilingual OCR system to make uni-lingual script helps us to reduce the requirement of two different OCR systems into a single OCR system for recognition of two different languages. This type of globalization helps the universal users of any language can read the text documents in their self-language if the bilingual documents are converted into uni-lingual document. In this paper, the image which contains printed Tamil and European numerals has been recognized using common OCR System and the Tamil numerals are converted into European numerals to globalize the document from a bilingual script into a uni-lingual document. The main objective of the work is to bring out the single numeral (European numerals) text document from the input image with two different numerals (Tamil and European Numerals). The Kohonen's self-organizing map (SOM) based recognition system has been used for recognizing the numerals and recognized characters in bilingual numerals (Tamil and European Numerals) form are converted into Uni-lingual form (European numerals). This paper also discusses the various approaches used for OCR.

Keywords: bilingual numerals recognition, segmentation, optical character recognition, Tamil OCR

1. Introduction

The development in the field of computers, digital documentation plays a vital role in all filed. The recognition of a bilingual character is a challenging job to the researchers. In this paper, the mixture of isolated Tamil and European numerals image is recognized using Kohonen's self-organizing map and converted into European numerals in order to globalize the bilingual document into uni-lingual document. The digital image of bilingual page has been scanned through optical scanners.

The scanned image is segmented using connected component method after preprocessing. The features of individual glyph of the image are extracted and then the SOM model is trained to recognize the text. The conversion process is followed by the recognition. The Tamil numerals are converted into European numeral. The final output is stored in a text document. The following Figure 1 shows the steps involved in the recognition and conversion process.

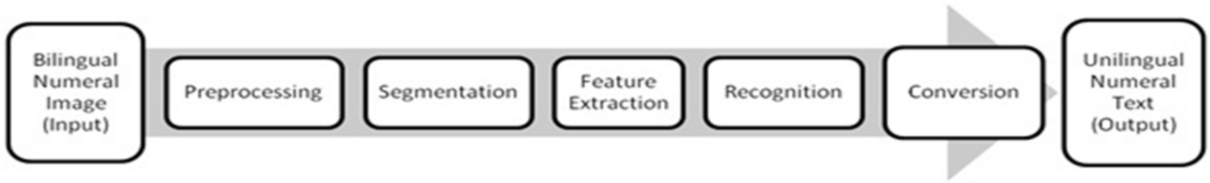
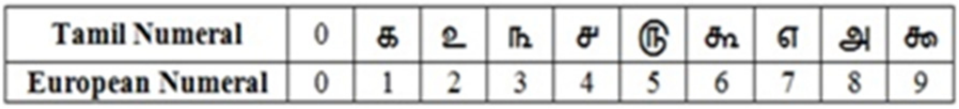


Figure 1. Uni-lingual OCR Steps

The following Figure 2 shows the equivalent European numerals for the Tamil numerals.



Tamil Numeral	0	௧	௨	௩	௪	௫	௬	௭	௮	௯
European Numeral	0	1	2	3	4	5	6	7	8	9

Figure 2. Equivalent European numerals for the Tamil numerals

Tamil numerals officially used by Tamilnadu, Sri Lanka, Singapore, Malaysia, Mauritius and other peoples around the world. But today European numerals which are used as common representations of numbers worldwide. Unlike European numerals, Tamil numerals have distinct representation for 10, 100 and 1000 as **௰, ௱ and ௲** respectively. In this paper, we considered the Tamil numerals from □ to ௯ which has an equivalent representation of European numerals as from 0 to 9.

2. Preprocessing

The preprocessing step involves Noise removal and binarization process which is useful for increasing the segmentation performance. The printed bilingual numeral image has been scanned through the optical scanners with the resolution of 100 dpi, and it is stored in the personal computer for further process. The input Bilingual image is shown in Figure 3 which is a scanned image. Initially, the RGB image is converted into a grayscale image. The grayscale image is binarized using Otsu's method (Otsu, N., 1978).

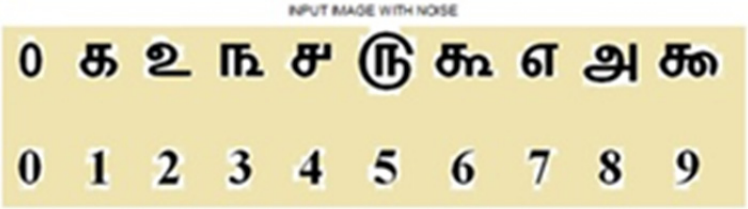


Figure 3. Bilingual Image Input

3. Segmentation

Segmentation is the process of obtaining the required multiple segments of the digital image.

Chen, D., Odobez, J.M. and Bourlard H (2004) have been introduced the edge detection method and morphological processing for text line detection and learning through Support vector machine for text rectangle identification. Eri Haneda, Charles and Bouman, A (2011) introduced the multiscale segmentation scheme for encoding of mixed raster content (MRC) document which is based on two algorithms namely cost optimized segmentation (COS) and connected component classification (CCC).

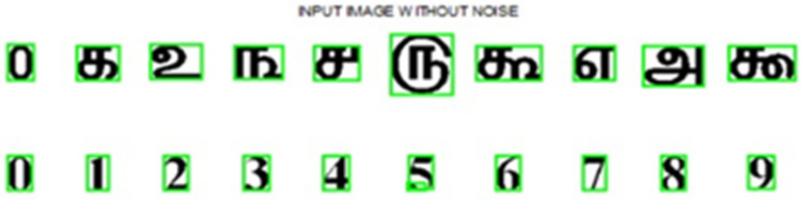


Figure 4. Binarized Image with Plotted Bounding Box

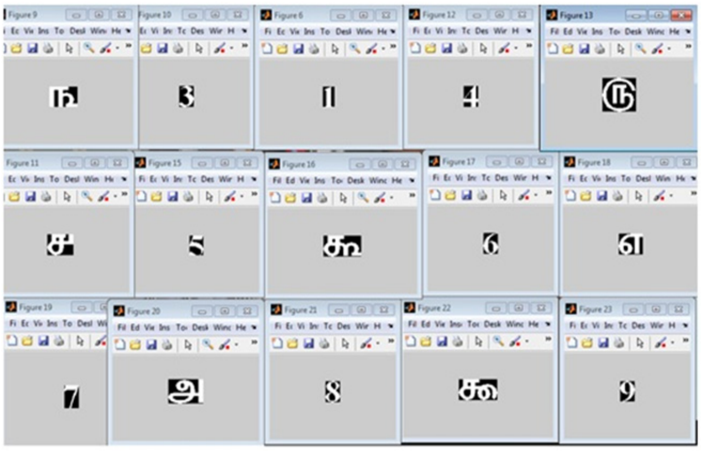


Figure 5. Few Segmented Character Images

In our work, the connected component analysis method has been used for detection of connected regions and bounding box has been plotted as shown in Figure 4. The selection of regions is done based on the properties of edge colour and line width of the preprocessed image. The snapshot of few segmented images of the bilingual input image has been shown in Figure 5.

4. Feature Extraction

Shivakumara. P, Weihua Huang and Tan, C.L. (2008) used statistical features like vertical and horizontal bars for segmented image edges for detecting the graphic and scene text. For Segmentation, the heuristic rules are created by grouping the filters and edge analysis. Xiangrong Chen and Yuille, A.L. (2004) have used AdaBoost learning algorithm to train the classifiers for recognizing the texts from natural scenes.

In this work, the individual image glyph features which help to achieve a higher recognition rate are measured. Each character attributes like number of horizontal lines, vertical lines, and the number of circles and arcs are defined for all the numerals used in recognition system. The following masks shown in Figure 6 and Figure 7 are used to define the attributes for detecting the horizontal and vertical lines.

-1	-1	-1
2	2	2
-1	-1	-1

Figure 6. Horizontal Mask

-1	2	-1
-1	2	-1
-1	2	-1

Figure 7. Vertical Mask

The horizontal and vertical masks are used to identify the presence of horizontal and vertical lines in the segmented image based on the gradient strength found.

The horizontal gradient component is calculated by, h[i][j] = f[i-1][j-1]*(-1) + f[i-1][j]*(-1) + f[i-1][j+1]*(-1) + f[i][j-1]*(2)+f[i][j]*(2) + f[i][j+1]*(2) + f[i+1][j-1]*(-1) + f[i+1][j]*(-1) + f[i+1][j+1]*(-1) (1)

The vertical gradient component is calculated by,

v[i][j] = [i-1][j-1]*(-1) + f[i-1][j]*(2) + f[i-1][j+1]*(-1) + f[i][j-1]*(-1) + f[i][j]*(2) + f[i][j+1]*(-1) + f[i+1][j-1]*(-1) + f[i+1][j]*(2) + f[i+1][j+1]*(-1) (2)

From the above equations, we obtained the gradient strength. The short and long horizontal lines are identified by setting up the threshold values. Similarly the features of vertical lines, arcs and circles are measured.

5. Recognition

The artificial neural network simplifies the recognition task and is extensively used in the pattern and character recognition (Mumtazimah Mohamad**,** Md Yazid Mohd Saman and Muhammad Suzuri Hitam, 2013). El-Yacoubi, A., Sabourin, R., Suen, C. Y. and Gilloux, M. (1999) used hidden Markov model-based approach for recognizing the handwritten words. Baum-Welch algorithm has been used for training and Viterbi algorithm logarithmic version carried out for recognition.

The unconstrained handwritten characters from single and multiple writer data are recognized using continuous density Hidden Markov models and Statistical Language model by Alessandro Vinciarelli, Samy Bengio and Horst Bunke (2004).

Dhandra, B.V., Gururaj Mukarambi and Mallikarjun Hangarge (2012) achieved recognition accuracy of 92.71% using *k*-Nearest Neighbor algorithm (KNN) and 96.00% using support vector machines (SVM)classifiers for handwritten Kannada vowels. In the same work, they have obtained 97.51% as recognition accuracy for KNN and 98.26% for SVM classifiers for handwritten English uppercase alphabets. Further, they obtained recognition accuracy of 95.77%, and 97.03% is obtained using k-NN and SVM classifiers respectively for mixed characters that contain Kannada vowels and English uppercase alphabets.

Sureshkumar, C. and Ravichandran, T., (2010) discussed about the recognition using SVM, RCS network, Self-organizing Map, Radial Basis Network and Fuzzy Neural Network for their extracted features of Tamil handwritten characters and achieved a higher recognition rate using RCS with the back propagation network. In this paper, Kohonen's self-organizing map (SOM) has been used for recognition task. The SOM is an artificial neural network trained using unsupervised learning method. The SOM uses a neighborhood function for preserving the input space topological properties. To determine the winner unit, the squared Euclidean distance between input and weight vector has been used.

5.1 Algorithm

Step 1: Initially the neighborhood parameters and learning rate are set. The weights are initialized. The input vectors are taken based on the attributes defined in feature extraction.

Step 2: For each input vector both Euclidean distance and weights are calculated.

Step 3: The Euclidean distance method is used to determine the winner unit after the input vectors are given. Squared Euclidean distance can be computed using the following expression.

$$D\ (j) = \sum (W_{ij}-X_i)^2 \quad (3)$$

Where

X=input vector i=1 to n & j=1 to m.

Step 4: The new weights are calculated as

$$W_{ij(new)} = W_{ij(old)} + \beta\ [\ X_i - W_{ij(old)}] \quad (4)$$

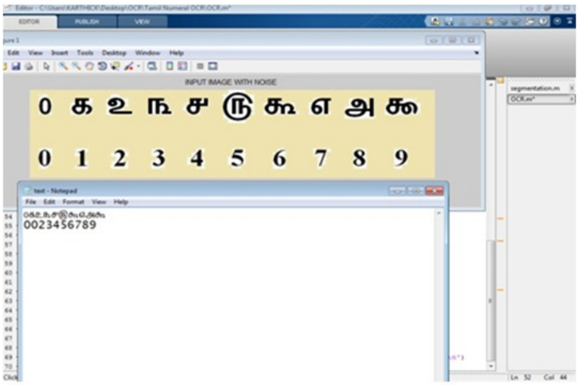
Where

β = Learning rate

Step 6: Learning rate β is updated.

Step 5: Radius of the topological neighbourhood has been reduced at specified times.

The Figure 8 shows recognized numerals for the given input bilingual image which is shown in Figure 3.



0 க உ ௩ ச ௫ ௬ எ அ ௯

0 1 2 3 4 5 6 7 8 9

```
0கஉ௩ச௫௬எஅ௯
0023456789
```

Figure 8. Recognized Numerals

6. Conversion of Bilingual document to Uni-lingual Document

The flowchart of the conversion process of the recognized characters is shown in Figure 9. Initially, the total number of characters has been counted. If the recognized character is a Tamil numeral then, the Tamil numeral is converted to equivalent European numeral which is shown in Figure 2. The conversion process is done by simple replacement of existing Tamil numeral. If the recognized numeral is a European numeral then, no replacement takes place.

The conversion process has been done until the character count becomes zero. The uni-lingual numeral output has been shown in Figure 10. Here all the Tamil & European numerals are in European numeral form.

7. Results

The input image which has printed numerals has been scanned through 'HP Scanjet 200 Flatbed Scanner' with the resolution of 100 dpi. It takes 14-15 Seconds to complete the scanning process. The personal computer of Intel Core 2 Duo CPU E700, 2.8 GHz processor with MATLAB R2013a version has been used to run the program. To train the self-organizing map, the learning rate is assumed as 0.3 and Initial Radius is assumed as zero. The Tamil numerals ௬ & ௯ are unable to recognize because of their similarities of assumed attributes. The various stages of results are shown in each section.

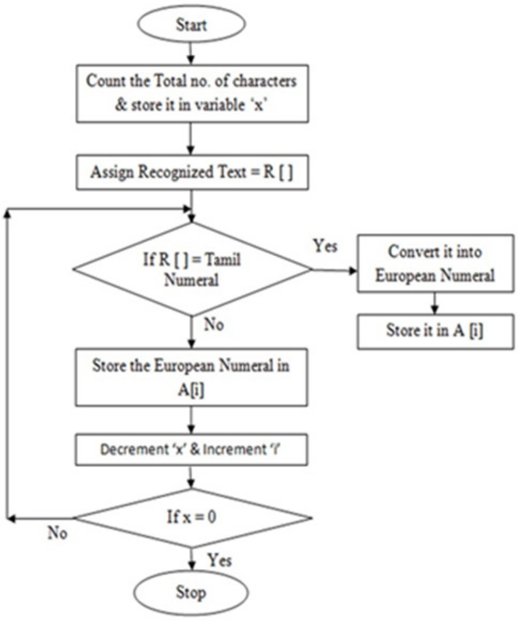
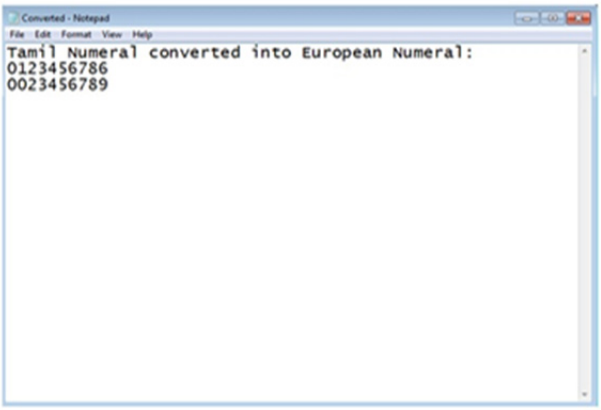

Start
Count the Total no. of characters & store it in variable 'x'
Assign Recognized Text = R [ ]
If R [ ] = Tamil Numeral
Yes
Convert it into European Numeral
Store it in A [i]
No
Store the European Numeral in A[i]
Decrement 'x' & Increment 'i'
If x = 0
No
Yes
Stop

Figure 9. Flow chart of Conversion Process



```
Tamil Numeral converted into European Numeral:
0123456786
0023456789
```

Figure 10. Uni-lingual Numeral Output

8. Conclusions

The bilingual OCR system to make uni-lingual numeral script has been created. The reduction in the requirement from two different OCR systems for recognizing two different language numerals to the single OCR system has been achieved by our single OCR system. This type of uni-lingual documentation may help for language translation to read the text documents in their self-language. In this paper, the image that contains printed Tamil, and European numerals has been recognized using common OCR System and the Tamil numerals are converted into European numerals. The input of two different numerals is brought into single numeral (European numerals) form. The Kohonen's self-organizing map based recognition system has been used for recognizing the numerals. The various approaches used for OCR also discussed. By using our system, the input of connected numerals can be recognized. In the future, the work can be extended for the rest of the Tamil numerals have distinct representation for 10, 100 and 1000 as ***௰, ௱ and ௲*** respectively.

References

Alessandro, V., Samy, B., & Horst, B. (2004). Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(6), 709-720.

Chen, D., Odobez, J. M., & Bourlard, H. (2004). Text detection and recognition in images and video frames. *Pattern Recognition*, *37*(3), 595–608.

Chen, X. G., & Yuille, A. L. (2004). Detecting and reading text in natural scenes. *Computer Vision and Pattern Recognition*, *CVPR 2004, Proceedings of the 2004 IEEE Computer Society Conference on, 2*, 366-373.

Dhandra, B. V., Gururaj, M., & Mallikarjun, H. (2012). Handwritten Kannada Vowels and English Character 12-17.

El-Yacoubi, A., Sabourin, R., Suen, C. Y., & Gilloux, M. (1999). An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition. *Journal IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(8), 752-760.

Eri Haneda, Charles, & Bouman, A. (2011). Text Segmentation for MRC Document Compression. *IEEE transactions on Image Processing*, *20*(6), 1611-1626.

Mumtazimah, M., Md Yazid Mohd Saman, & Muhammad, S. H. (2013). Divide and Conquer Approach in Reducing ANN Training Time for Small and Large Data. *Journal of Applied Sciences, 13*(1), 133-139.

Otsu, N. (1978). A threshold selection method from grayscale histograms. *IEEE transactions on Systems, Man, and Cybernetics*, *8*(1978), 62-66.

Shivakumara, P., Weihua, H., & Tan, C. L. (2008). An Efficient Edge Based Technique for Text Detection in Video Frames. *The Eighth IAPR International Workshop on Document Analysis Systems*, 307-314, DAS '08. 16-19 Sept. 2008.

Sureshkumar, C. & Ravichandran, T. (2010). Handwritten Tamil Character Recognition and Conversion using Neural Network. *International Journal on Computer Science and Engineering (IJCSE)*, 2(7), 2261-2267.

Copyrights