Script Identification Based on Morphological Reconstruction in Document Images

B.V. Dhandra, P. Nagabhushan, Mallikarjun Hangarge, Ravindra Hegadi, V.S. Malemath
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga-585106, Karnataka, India
dhandra_b_v@yahoo.co.in, pn@amrita.edu, mhangarge@yahoo.com*

Abstract

In this paper, the study of script identification based on morphological reconstruction for printed document images is carried out. The system is developed by using 609-scanned document images representing English, Hindi, Kannada, and Urdu scripts. The system developed includes a feature extractor and a classifier. The feature extractor consists of two stages. In the first stage, the morphological erosion and opening by reconstruction is carried out on a document image in horizontal, vertical, right and left diagonal directions using the line structuring element. The length of the structuring element is fixed, based on the average height of all the connected components of an image. In the next stage, average pixel distribution is found in these resulting images. A nearest neighbor analysis is used to classify the new documents. Accuracy of classification averaged 97% across the four scripts. The method shows robustness with respect to noise, font sizes and styles.

1. Introduction

With recent emergence and widespread application of multimedia technologies, there is an increasing demand to create a paperless environment. Hence, document image processing in general and optical character recognition (OCR) in particular will play an important role in transformation of the traditional paper based environment to truly paperless electronic environment.

The OCR is of special significance for a multi-lingual country like India. Although a large number of OCR techniques has been developed over the years [10, 12], almost all-existing works on OCR assume that the script and language of the document is known beforehand. Thus, individual OCR tools have been developed to deal with only one specific language [1]; i.e., an OCR developed for English will work satisfactorily for English document with desired accuracy, whereas this system not at all work satisfactorily for the documents like Kannada, Hindi, etc. In view of this, it is essential to develop a pre-OCR script and language identification system for multi-script and multi-lingual country like India. This has motivated us to propose a method for automatic script or language identification from document images containing any of these proposed scripts.


In this paper, we have extended the work [8] by considering scanned document images of four scripts/languages with different font sizes and styles by relaxing the font size and style constraint of [8].
This paper is organized as follows. The section 2 describes some discriminating features in the characters of Kannada, Hindi, English and Urdu. The section 3 describes a proposed new method for identifying the four scripts- Kannada, English, Hindi and Urdu. The section 4 describes proposed algorithm. The experimental details and the results obtained are presented in section 5. Conclusion is given in section 6.

2. Discriminating features in the characters of Kannada, Hindi, Urdu and English text

The feature extraction is the integral part of the any recognition system. The aim of feature extraction is to identify patterns by means of minimum number of features that are effective in discriminating pattern classes. The new algorithm is inspired by a simple observation that the presence of vertical, horizontal, right, left diagonal strokes and holes in the characters of the four proposed scripts are more distinct. Most of the Hindi (Devnagari) language characters have horizontal (sirorekha in Devnagari [7]) and vertical stokes like structure [9]. It has been found that a distinct property of the English characters is the existence of the vertical strokes like structure [7]. It could be seen that most of the Kannada characters have horizontal stokes like structures [9] and also strokes in right and left diagonal directions, whereas Urdu characters contains more strokes in right diagonal and horizontal directions. The Urdu characters have less number of holes compared to English, Kannada and Hindi. Hence, these directional stroke features are considered to be more potential to distinguish each script.

3. The Proposed Method

The proposed method is based on five discriminating features such as (i) the characters that contain horizontal strokes (ii) vertical strokes (iii) right diagonal strokes (iv) left diagonal stokes and (v) the characters that contain the holes in the document image.

3.1. Pre-processing

The images are scanned using HP Scanner at 300 DPI, which usually yields a low noise and good quality document image. The digitized images are in gray tone and we have used Otsu’s global thresholding approach to convert them into two-tone images The two-tone images are then converted into 0-1 labels where the label 1 represents the object and 0 represents the background. The next step in pre-processing is skew detection and correction. However, we assume that the skew correction has been performed before. We have applied thinning operation only once on input image and removed the lighter objects touching the border of an image by morphological reconstruction. We have also used morphological opening to remove small objects like speckles from binary image to extract more distinct features. Eight neighbors connectivity is used for this operation.

3.2 Reconstruction and Hole Filling

The algorithm computes the eight-connected components of white pixels on the image and produces the bounding boxes for each of the connected components. The average height of the bounding boxes is used to fix the length of the structuring element for morphological erosion and opening operations. To extract the characters containing strokes in four directions (Horizontal, vertical, right diagonal and left diagonal), we have performed the erosion operation on the input binary image with the line-structuring element. The length of the structuring element is thresholded to 70 % (experimentally fixed) of average height of all the connected components of an image. The erode image is used for opening by reconstruction in the vertical, horizontal, right and left diagonal directions using a fast hybrid reconstruction algorithm [2]. The reconstructed images of four scripts are illustrated in figure 1. The reconstruction is a morphological transformation involving two images and structuring element. One image, the marker, is the starting point for the transformation. The other image, the mask, constrains the transformation. Here, the structuring element used defines the 8-neighbors connectivity. We have used erode image as the marker and original image as the mask throughout the experiment for reconstruction. The reconstructed images (in four directions) and its original image are used for fill holes. For fill holes, we choose the marker image (erode image), f_m, to be 0 everywhere except on the image border, where it is set to 1-f. Here f is the original image.

\[ f_m(x, y) = \begin{cases} 1-f(x, y), & \text{if } (x, y) \text{ is on the border of } f \\ 0, & \text{Otherwise} \end{cases} \]

Then \( g = [R_c (f_m)]^c \) has the effect of the filling the holes in \( f \) as shown in figure 1. Where, \( R_c \) is the reconstructed image. Finally, the number of on pixels left in the images (reconstructed with fill holes) in figure 1 is counted. Then a feature value is defined as the number of on pixels divided by the total number of pixels.
pixels in an input image. Thus a set of five features is obtained for each document image. For classification, Euclidian distance is calculated between test and training feature vector and based on the minimum distance criteria classification of four scripts is carried out.

4. Proposed Algorithm

The different steps involved in the proposed algorithm are as follows.
1. Pre-process the input document image i.e. Binarisation using Otsu’s method, apply thinning and remove the border objects and speckles using Morphological reconstruction.
2. Compute the number of connected components of the pre-processed document image.
3. Carry out the Morphological erosion and opening by reconstruction using the line structuring element in vertical, horizontal, right and left diagonal directions.
4. Fill up the holes of four reconstructed images of step 3 and its original image.
5. Compute the average pixel distribution of resulting images of step 4 and which produces a five features vector for each document.
6. Compute Euclidean distance between test and training vector obtained in step 5. Classify four scripts based on minimum distance nearest neighbor classifier.

5. Results and Discussion

We have applied the proposed script identification algorithm on two data sets. In the first set, we have considered 400 document images. These images are scanned by a flatbed HP scanner at a resolution 300 dpi. The size of the sample image considered is 128x128 pixels. The first set document images are scanned from newspapers, magazines, weeklies, and schoolbooks containing variable font’s, styles and sizes. Out of these 400 documents images English, Hindi, Kannada, and Urdu are 100 each. From the experiment on the above data we noted that overall accuracy of the system was about 97%. The distribution of the accuracy achieved in identifying the scripts is presented in Table 1. Although the primary aim of this paper is achieved, that is the script identification based on morphological reconstruction approach; it is a fact that, normally printed documents font sizes and styles are less varied. We therefore, conducted a second set of experiment on 209 document images to test the sensitivity of the algorithm towards different font sizes and styles. These images are first formatted in different fonts using DTP packages, and then printed from a laser printer. The printed documents are scanned as mentioned earlier. The ISM, DTP package is used for Hindi and Kannada, Microsoft word for English. On most commonly used seven fonts of Kannada, Hindi and five fonts of English scripts are considered for experiment. For each font 11 images are considered varying in font size from 10 to 36. Out of these 209 document images, 55-English, 77-Kannada, and 77-Hindi. Further script identification accuracy achieved for Kannada script is 100% with respect to the fonts KN-TTKaveri, TTNandi, TTPadmini, TTPampa, TTUma, TTRanna and 90% with TTradhey. The Classification accuracy of Hindi script is 100% for the fonts, DV-TTBhima, TTRadhika, TTAakash, TTNatraj, TTsurekh, TT Vasundhra, and TTYogesh. Similarly, English script identification result is 100% for the fonts Tahoma, Arial, Bookman Old Style, Arial Narrow and 90% with Times New Roman. The classification results of second data set reveals better than the first data set. This is because of paper and print quality. From the experiment on first data set we noticed the overall error rate of 3% of the proposed system. Further 5% with respect to Kannada script is due to the poor quality documents. 3% with respect to English is due to double line spaced documents. By normalizing the text [1] the error rate can be minimized with respect to English. Here we have not attempted the text normalization. The 2% error with Hindi and Urdu is due to the documents containing broken characters. The proposed algorithm is implemented in MATLAB 6.1. The average time taken to recognize the script of a given (128x128) image is 1.2531 seconds on a Pentium-IV based machine running at 1.80 GHz. with 128 MB RAM. We have applied the same algorithm without thinning operation on 215 handwritten document images collected from 100 writers in English, Hindi and Kannada scripts. The classification accuracy is 90.67 %, 84.29% and 81.42% with respect to Kannada, English and Hindi. Overall accuracy is 85.46%. Our algorithm is simple, robust and reliable as compared to [8].

6. Conclusion

To make a successful multi-script OCR, it is necessary to identify different scripts before feeding to OCR system of individual script. Thus, the work presented here has a strong direct application potential. The morphological reconstruction approach is the first of its kind for script identification. The novelty of this method is font size and style independent.

7. Acknowledgement

The authors are grateful to Dr. Umapada Pal, ISI, Kolkata, Dr. G. Hemanthkumar and Dr. D.S. Guru,
Table 1. Script identification results

<table>
<thead>
<tr>
<th>Script</th>
<th>Kannada</th>
<th>English</th>
<th>Hindi</th>
<th>Urdu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kannada</td>
<td>95%</td>
<td>02%</td>
<td>02%</td>
<td>01%</td>
</tr>
<tr>
<td>English</td>
<td>01%</td>
<td>97%</td>
<td>02%</td>
<td>02%</td>
</tr>
<tr>
<td>Hindi</td>
<td>01%</td>
<td>01%</td>
<td>98%</td>
<td>0%</td>
</tr>
<tr>
<td>Urdu</td>
<td>01%</td>
<td>01%</td>
<td>0%</td>
<td>98%</td>
</tr>
</tbody>
</table>

Figure 1. Images (a) to (d) represent vertical, (e) to (h) horizontal, (i) to (l) right diagonal, (m) to (p) left diagonal reconstructed images with fill holes of English, Hindi, Kannada, and Urdu respectively, images (q) to (t) represent original images with fill holes of English, Hindi, Kannada, and Urdu respectively.