

Template matching approach for Handwritten Kannada Numeral Recognition

Vishweshwarayya C Hallur

Department of Computer Science,
Karnataka University, Dharwad, INDIA

Ravindra S. Hegadi

School of Computational Sciences,
Solapur University, Solapur, INDIA

ABSTRACT

In this paper we propose a simple template matching method based on correlation co-efficient to recognize the hand written Kannada numerals. In this, hand written Kannada numeral documents are scanned to preprocessors to eliminate noise. By applying line segmentation each numeral is extracted. And then each segmented object is calculated between the each numeral under consideration and stored numeral image data. A high correlation coefficient will indicate the successful match between the numeral under consideration and stored numeral.

Keywords

Kannada Numerals, Template matching.

1. INTRODUCTION

The problem of Kannada numeral recognition is a part of characters recognition, which is part of document image analysis. In a hand written image document analysis the document image is preprocessed for extraction of information needed. Optical characters recognition (OCR) is an important task carried out on text reorganization [1]. There are many commercial OCR are available for English, Chinese, Japanese and most of the European languages with good accuracy. Only few works are reported for Indian languages, the accuracy claimed by these works is not quite reasonable [2], [3]. This is due to the reason that Indian languages are very complicated in their structure and also there are more number of literals in Indian languages as compared with English. We can find different feature extractions methods in literature for recognition of numerals belonging to different Indian scripts. Pal and B. B. Choudhari [4] have reported on complete printed Bengal OCR system for recognition of Bengal numerals and characters. Recognition of hand written Devanagari numerals is described in [5]. In this work three different types of features are used, namely density feature, descriptive component features and moment features. In [6], fuzzy based approach is proposed for the recognition of multi-font numerals. In [7] multi-font numeral recognition without tinning based on directional density of pixels is reported. A 10 - segment string concept is reported in [8], which is based on water reservoir principle, horizontal and vertical strokes and end points. These are used as features for Kannada hand written numerals recognition. In [9] G.G. Rajapat has reported on image fusion and basic nearest neighbor classifier recognition of Kannada hand written numerals. In this paper, a template matching method is proposed using correlation coefficient computation for recognition of hand written Kannada numerals. In section 2, the methodology of proposed work is discussed. In section 3, the experimentation is carried out in hand written Kannada numerals and in section 4, conclusion are drawn.

2. PROPOSED METHOD

The hand written documents are scanned and those Images are preprocessed to eliminate noise and then resized to a standard

size. Each image template is matched with the stored image data from the database. The detail description of this process is given in the following sub sections.

2.1 Image processing

Like other languages Kannada language has 10 basic numerals as shown in figure 1. Thresholding is applied to convert scanned Kannada numeral document image to binary image. The noises in the images like tiny dots are eliminated by removing all connected regions having less than 30 pixels. There are multiple lines in the preprocessed image, each line containing multiple numerals. The document is scanned when pixel is found from top-left corner of the image and its corresponding bottom-right pixel is located. With the help of these pixels, the rectangle is formed and it is extracted, which will be one line of text containing numerals approach, may require searching potentially large amounts of pixels in order to determine the best matching location [7]. This approach uses the entire template, with generally a sum-computing matrix which determines the best location by testing all or a sample of the viable test locations within the search image that the template image match up to. A template based approach may be useful for templates without strong features. In this numeral recognition whole template is matched with each segmented numeral image. Thus the proposed method is one of the cases of template based approach. Template based matching may require sampling of large number of pixels of certain image processing problems. It is possible to reduce the number of sampling points by reducing the resolution of the searching image and template images by the same factor and then performing the operation on the downsized images, providing search window of data points within the search image so that the template does not have to search every viable pixel or combination of both. A training data set for each numeral is created and stored in the database. Here hand written numerals are taken as training data set. Again the hand written numeral in the training data set will have the size of 42×24 . For each character a bounding box is drawn and coefficient is calculated between the segmented character from the input image and every character in the database. The correlation coefficient is a real value between -1 and 1. If there is no relationship between numeral image under consideration values and the stored image which is matched, then the correlation coefficient is 0 or very low. As the strength of the relationship between the predicted values and the actual values increases so does the correlation coefficient. There may be many such lines in the in the image document. The extracted line segmented into individual numeral image by identifying the connected components and drawing the bounding box for each connected component. Each segmented numeral image will have different spatial resolution. Before further processing, all segmented characters are resized to a standard size of 42×24 pixels.



Fig.1. Printed Kannada numerals.

2.2 Template matching

Template matching is a technique in a Digital image processing for finding small parts of an image which match template image. Template matching can be divided into two approaches: feature based approach and template based approach matching. In the feature based approach which uses the features of the search and template image, such as exist of an image as a primary measuring metrics to find the best matching location of the template in the source image. When image has strong features then by that time the feature based template matching may be considered. Since feature based approach does not consider the complete template image, it can be more computationally efficient when working with source images of larger resolution. As the alternative approach, i.e. template based the predicted values and actual values increases so does the correlation coefficient. A perfect fit gives a coefficient of 1.0. The mathematical formula for computing r is

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of pairs of data, x and y are the training and testing numeral image data. More will be the match between the numeral with the template from the database for the higher value of r . For each numeral from input image ten correlation coefficient values are generated from 0 to 9. The maximum value among these will show the matching numeral from the database.



Fig.2. input image containing Kannada numerals.

3. EXPERIMENTAL RESULTS

The input fonts generated include the hand written fonts for implementation Matlab version 10 software is used. Figure 2 shows an input scanned image containing Kannada numerals from 0 to 9 in a row. This input scanned image is preprocessed by eliminating the noise. Noise is eliminated by identifying the entire connected regions in the input image with less than 30 pixels, which corresponds to the noise. The preprocessed image is segmented line by line by scanning and locating the left-top edge and right-bottom edge of the line from the document as shown in Figure 3. It can be noticed that this segmentation will extract a line of numerals. Where all sides of these lines of numerals will exactly touch the boundary of the bounding box. The next task will be to segment on individual numerals from each segmented line. This process will draw bounding box for each numeral which

touches all the four sides of the numeral. Such segmented numeral images are resized to 42×24 pixels.

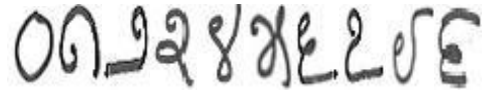


Fig.3. segmented numeral images.

4. CONCLUSION

In this paper, template matching approach is proposed to recognize handwritten Kannada numerals. Template matching algorithm is old and well known method. There is an attempt made to implement this algorithm based on correlation coefficient to recognize and classify the handwritten numerals. The most of the Kannada numerals are formed with curves. The results of proposed method are quite encouraging. The performance of the proposed algorithm can be improved by considering more features like topological properties of Kannada numerals, curvatures, etc.

5. REFERENCES

- [1] Gorman L.O., Kasturi R.: Document image analysis, IEEE Computer Society Press (1995).
- [2] Choudary B.B., Pal U.: An OCR system to read to Indian languages scripts: Bangla and Devanagari. In: Fourth Int. Conf. Doc. Ana. Rec., PP. 1011-1015 (1997).
- [3] Sinha R.M.K., Mahabala H.: Machine recognition of Devanagari script. IEEE Trans. Sys., Man, Cybern. q, pp. 435-449 (1979).
- [4] Kunte R. S., Samuel R.D.S.: An OCR system for printed Kannada text using two-stage Multi-network classification approach employing Wavelet features. In: Int. Conf. Computational Intel. Multimedia Appl., pp.349-455 (2007).
- [5] Ashwin T.V., Sastry P. S.: A font and size independent OCR system for printed Kannada Documents using support vector machines. Sadhana 27(1), 35-58 (2002).
- [6] Rajaput G.G., Hangarge M.: Recognition of isolated hand written Kannada numerals based on image fusion method. In: Proc. PReMI, pp.153-160 (2007).
- [7] Yuhai L., Jian L., Jinwen T., Honbo X.: A fast rotated template matching based on point feature In: Proc. SPIE 6043, pp. 453-459 (2005).
- [8] Dinesh Acharya U, N V Subbareddy and Krishnmoorthy, Isolated Kannada Numeral Recognition Using Srtuctural Features and K-Means Cluster, Proc. Of IISN-2007, (2007) 125-129.
- [9] G.G.Rajaput and Mallikarjun Hangarge, “Recognition of Isolated Kannada Numeral Based on Image Fusion Method”, PReMI 2007, LNCS 4815, pp.153-160, 2007.