Word Spotting from Handwritten Kannada and Roman Text Documents

Veershetty C

Department of Computer Science, Government First Grade College, Basavakalyan-585327, India

Abstract: ***In this paper, we reported the efficient method for word spotting from handwritten Kannada and Roman text documents. Our main aim is to classify the given word whether it's Kannada or English based on keyword. Firstly, words are extracted from scanned document images. Then Radon transform with different orientation and texture features such as local entropy, local range, and local standard deviation are used as features from each word image. Popular Euclidean distance is used for classification of the document and obtained average 99.25% accuracy in word spotting form document images.***

Keywords: Handwritten Script, Kannada, Roman, Radon Transform, Cosine, OCR

1. Introduction

The world is becoming automated and multilingual with the advancement of digital profession. Digital solutions are provides easy storage, access and retrieval of information which stored digitally. Increasing the need of paperless office has given a huge space to research in document image analysis and recognition. In automatic document image processing method Optical Character Recognition (OCR) is the key research area which is playing very important role in office automation. Optical Character recognition is the process in which paper document is converted into digital form with the help of digitizer and then converted in to machine editable format.

Unfortunately OCR methods are designed with the assumption that document which is going to be process containing known script. In such case, script dependent OCR is not feasible to process documents in multilingual environment. This problem can be handled by designing Multilingual Optical Character Recognition (MOCR). Which is very difficult and challenging task for the country like India, which having 12 scripts and 22 official languages. Therefore the ultimate solution to overcome this problem is the use of spectrum of OCR for different scripts. However this needs the identification of script from document image and channelize towards its particular OCR system.

In our daily life, we come across with many documents which consisting of multiple handwritten text, like bank cheque images includes amount in handwritten, postal envelopes with handwritten addresses, handwritten information filled in application form etc. Automatic processing of such documents is also needed to identify the script of handwritten text, because in India most of documents are bilingual or trilingual in nature (Roman script with regional language and Roman, Hindi and Regional Language). Identifying handwritten script is very challenging task as compared to printed one. Different font styles for a script in printed document have similar kind of structure, but in handwritten document there is huge variation in writing styles. Therefore problem of handwritten script identification becomes more difficult. In this paper, we have made an attempt to identify the script of handwritten words written in Kannada and Roman text.

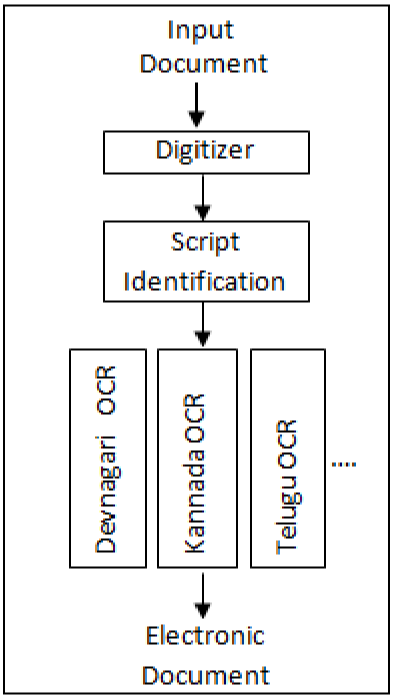


Figure 1: Multiscript OCR farme work

2. Related Work

In the past research, state-of-the art algorithms are presented in literature which concentrates on script identification in Indian handwritten document images. Among them [1] proposed an automatic scheme for word-wise identification of hand-written scripts for Indian postal automaton. In the proposed scheme, features like, water reservoir concept based features, fractal dimension based features, topological features, and scripts characteristics based features etc. are given to neural network classifier for classification of roman and Oriya script. In [2] DCT and Wavelet features are employed for script identification. Block level trilingual script classification is reported in a fashion that each regional Indian script has given opportunity to be with Roman and Devanagari Script. In [3] Gabor filters are used for feature computation with K- NN classifier for line level script identification. In [4] reported a scheme for script identification on line level and block level .13 different spatial features based on morphological operations are calculated and k-NN classifier is employed for classification. In [5] designed a Multi Layer Perception (MLP) based classifier for script separation, Where 8 different word-level holistic features used to train the classifier.

Literature review shows that there are only some works are reported in the context of handwritten script identification, which carried out on different scripts on different levels such as block level, line level and word level. Level of script identification is decided with respect to the different applications, in our context i.e. Multilingual Handwritten Character Recognition (MHCR), we chosen word level identification of script.

In section I and II we have discussed our motivation and past work. In section III we have given detail description of proposed approach. Section IV devoted for experiments and analysis. We have concluded in section V.

3. Proposed Approach

We intended to identify the script from Kannada and Roman handwritten text words. To do this task, first words are extracted from handwritten document images. For each extracted word Radon transform based eight features are computed with texture Features such as local standard deviation, local entropy and local range, these eleven features are used to from a feature vector, which is then used for script identification using k-NN classifier. In Fig 2. We have shown the work flow of our proposed approach.

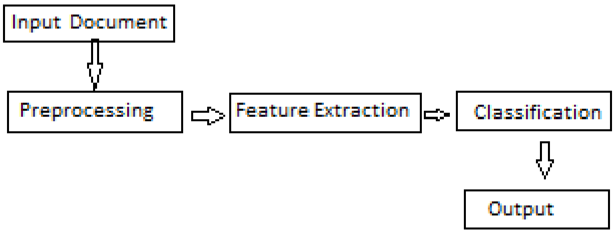
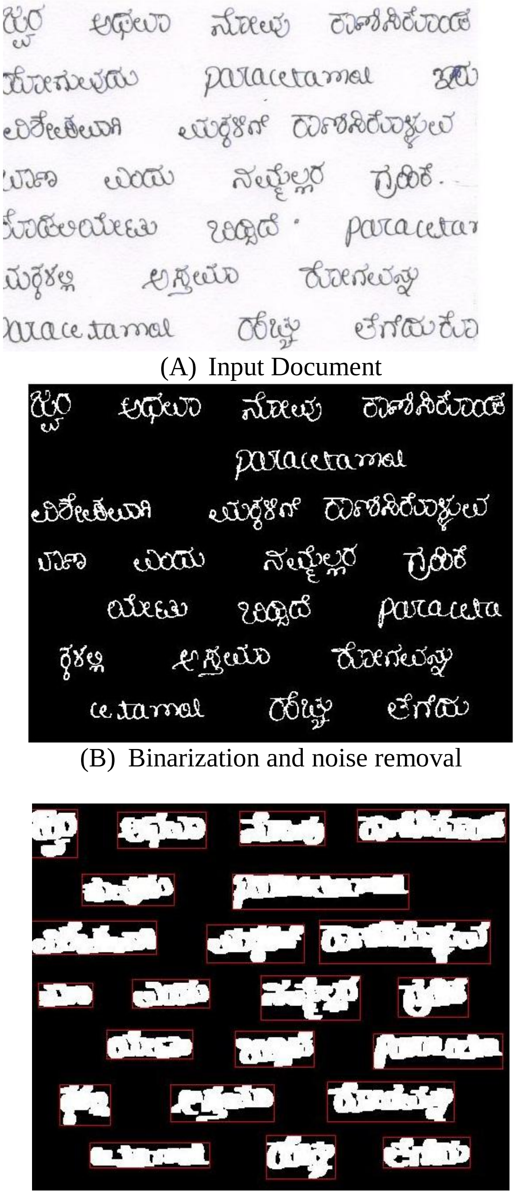


Figure 1: Screen shot of proposed approach

*a) **Preprocessing***

First step in our approach is to do preprocessing. Preprocessing is done with the aim of word extraction from document images. It involves three stages, in first stage document is binarized using Otsu's threshold selection method. In second stage morphological operators are employed for removing commas, quotation marks, full stop etc. In third stage morphological dilation is applied and using simple connected component rule words are segmented. Visualization of segmentation process is shown in Fig 3.


(A) Input Document

(B) Binarization and noise removal

(C) Dilation and word extraction

Figure 2: Pre-processing and word segmentation

*b) **Feature Extraction***

To compute the features we used radon transform and texture features such local entropy, local range, local standard deviation. Radon transform is frequency based technique which computes the projection of image matrix towards a particular direction. A projection of two dimensional function **f (m, n)** is a set of line integrals. Radon transform computes the line integral from multiple sources along parallel paths in the defined direction. Then the resulting projection is sum of the energy of the pixels in each direction, which is line integral. Resultant image will be **R (ρ, θ),** which is given by:

$$\rho = x \cos\theta + y \sin\theta \qquad (1)$$

Radon transform given as follows

$$R(\boldsymbol{\rho},\boldsymbol{\theta}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\delta(\rho - x\cos\theta - y\sin\theta)\,dx\,dy \qquad (2)$$

Where δ is Dirac delta function.

We have computed 1 to 8 features from radon transform of word image with the direction of θ= {0, 30, 65, 85, 130, 155, 240, and 250}. These directions are decided experimentally. Illustration of Radon transform is shown in Fig. 4.

Further 9 to 11 features are computed using Statistical Filters. Statistical filters are aimed to measure the information about local variability of the energy of pixels in an image and we used three statistical filters namely Mean filter, Entropy filter and Std Filter. Statistical Filter returns the image g(x; y) where each output pixel contains the statistical value (mean, std & entropy) of the 3x3 neighborhood around the corresponding pixel from input image f(x,y), as a feature we used standard deviation of these filtered images. Visualization of entropy filter is shown in Fig. 5.

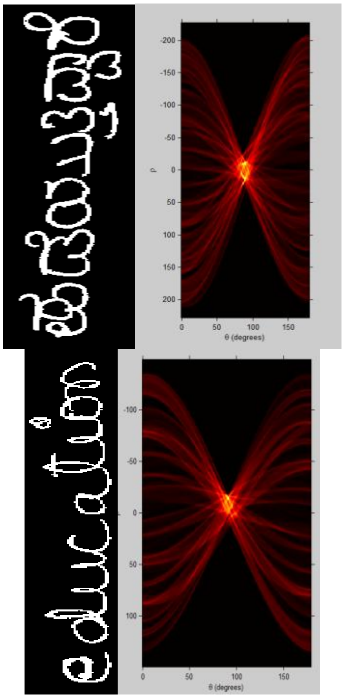


Figure 3: Handwritten word and their corresponding Radon Transform

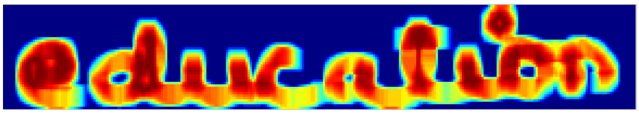


Figure 4: Roman Handwritten word image filterd with Entropy Filter

c) *Classification*

Most popular and conventional technique Nearest Neighbor Classifier is used for classification. It is supervised learning algorithm.

The nearest neighbor can be searched using suitable distance measure. In this work we use Euclidean distance as distance measure to find nearest neighbor. Let P(p1,p2,p3…pn) be the training data and Q(q1,q2,q3,q4…qn) be the testing sample , the Euclidean distance between P and Q is given as follows:

$$D(P,Q) = \sqrt{\sum_{i=1}^{n} (P_i - Q_i)^2} \qquad (3)$$

4. Experiments

1) Data Set

There is no standard dataset is available for the handwritten script identification in bilingual Kannada and English document images. In this paper, we collected sample of 50 writers on A4 size pages. By applying segmentation algorithm described in section 2.1 we extracted word images, correctly segmented 1000 word images of Kannada script are chosen for experimentation. For Roman scripts IAM DB 3.0 [7] is utilized and obtained 1000 word images. In this way dataset of 2000 word images of Kannada and Roman script is created. Sample words from our dataset can be seen is Fig.6.

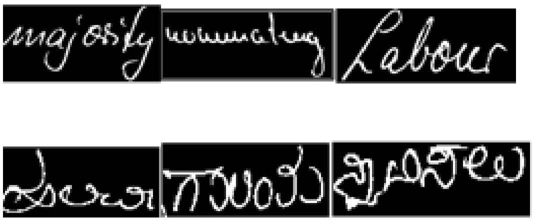


Figure 5: Roman Handwritten word image filterd with Entropy Filter

2) Evaluation Protocol

Most popular method is used for evaluation of proposed approach, named as k- fold cross validation. The dataset is randomly partitioned in to k sub folds. Each time one fold is consider for testing and rest for training. Process is repeated until each fold has got opportunity to serve for training and testing. In our experiment we considered k= 10. We have defined accuracy as following:

$$\text{Accuracy} = \frac{\text{\#Correctly Classified Words in the Class}}{\text{\#Total Words in the Class}} \text{ X } 100$$

3) Results

To test discriminative power of our algorithm we have carried out exhaustive experiments with 2000 word images of our dataset. It consist 1000 word images for Kannada and 1000 word images for Roman script. Our exhaustive experimental tests are goes like this:

1) Compute the feature vector of fixed size from each isolated word
2) Perform the nearest neighbor classification with Euclidean distance
3) Evaluate the method using 10 fold cross validation

In Table I we have given the results for Identification of handwritten Kannada and Roman words. Identification for Kannada script is 99.30% and for Roman we noted 99.20 % accuracy. Results are speaking that proposed method is potential to identify handwritten scripts in bi-script handwritten documents. For deeper analysis we have given confusion matrix for Kannada and Roman handwritten text words in Table II. We have observed that misclassification occurred due to very small size of words, for example "a" in Roman script. Due to the unavailability of common benchmark data set it's not possible to do the fair comparison with previous work.

Table 1: Script Identification results in accuracy

Script	Identification Accuracy
Kannada	99.30%
Roman	99.20%
Average	99.25%

Table 2: Confusion Matrix

Script	Kannada	Roman
Kannada	993	7
Roman	8	992

5. Conclusion

In this paper, we presented the study of Radon transform and texture features for identification of script from handwritten Kannada and Roman text words. It can be noted that our algorithm is not much sensitive to writing style, ink and size of text words. It is partially invariant to skew; it can be noted from Fig 6.

Further we dedicated to develop the state-of-the-art algorithm for script identification in Indian multi script handwritten document images

References

[1] K. Roy, U. Pal, "Word-wise Handwritten Script Separation for Indian Postal automation", in Proc. of IWFHR, La Baule, France, Oct, 2006.

[2] G G Rajput and Anith H B., 2010. Handwritten script recognition using DCT and wavelet features at block level. Recent trends in image processing and pattern recognition, pp 158 – 163

[3] G G Rajput , Anita H B , " Handwritten Script Identification from a Bi-Script Document at Line Level using Gabor Filters" in Proc. Of Proceedings of the International Workshop on Soft Computing Applications and Knowledge Discovery (SCAKD 2011), pp -94-101.

[4] Mallikarjun hangare and B V Dhandra., 2010. Offline handwritten script identification in document images. Intenational journal of computer applications.Vol 4, pp 6 – 10.

[5] Ram sarkar,Nibaran Das, Subhadip Basu, MahantapasKundu., 2010.Word level script identification from bangla and devanagari handwritten texts mixed with Roman script. Journal of computing, Vol 2 , pp 103 – 108.

[6] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.

[7] U V Marti, H Bunke, The IAM Database :an English sentence database for handwiritn recogntion , International Journal on Document Analysis and Recognition Volume 5, Issue 1(2002) , pp 39-46