International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

Word-Wise Tri-Lingual Script identification using K-NN and SVM

Renuka Devi B[1], Raghavendra Srinivas[2]

*[1]Computer Science Department, JSS Manjunatheshwara Institute ofvUnder-Graduate and Post-Graduate Studies,Vidyagiri, Dharwad, Karnataka, India

*[2]Department of Computer Science, University of Horticultural Sciences, Bagalkot, Karnataka, India

ABSTRACT

This paper presents the script identification for tri-lingual based on K-NN and SVM classifier. For the proposed three languages were utilized namely: Kannada, Hindi and English. For the experiment 6000 word images dataset has been used it includes 2000 images belongs to each language. For the features LBP features are extracted from word images. The no. of features are 59 obtained from LBP method. For the recognition K-NN and SVM classifier has been used accuracy. The optimum result is for K-NN is 98.38% and for SVM 98.50% are obtained.

Keywords: Script identification, word wise images, Document Image, K-NN, SVM.

I. INTRODUCTION

The identification of scripts is treated as a important step in automatic processing of multi-lingual documents. This step is helpful to choose proper OCR for proper language for the process. The proposed method is addressed the problem script identification in the tri-lingual scripts. In current scenario there is a much increasing the text document generation, which are commonly bi-lingual or tri-lingual scripts. If the document has the bi-lingual; it may contain regional language with English language, if the document contains the tri-lingual that should includes regional, national and international (English). This kind of document may need for further processes. The proposed work taken the three language scripts namely Kannada, Hindi and English these scripts is used in the Karnataka state, India.

This method is has been used the printed words standard dataset[1] out of 11 languages word images we have considered Kannada, English and Hindi word dataset. To extract features the LBP feature extraction method has been employed, the method has given good discriminative features.

For the classification purposed the proposed method is employed K-NN and SVM classifier which are supervised learning methods. For the K-NN Euclidian distance metric used and for SVM Cubic Kernel function is used and both the classifiers are given promising results of 98.2% and 98.5% respectively.

II. Literature Survey

We found some paper in relation with the script identification, Peeta Basa Pati et al. [9] has proposed script identification based on HVS based system for script identification. Peeta Basa Pati et al. [1] has addressed the word level multi-script identification. T.N.Tan[5], implemented an automatic script identification by using rotation invariant texture features. D.Dhanya et al.[10], worked based on Gabor filter and spatial spread filter features. A.Busch et al.[7] addressed the work on script identification based on texture. Judith Hochberg et al.[3] proposed a work on handwritten script and language identification.

III. Data Collection and Preprocessing

To implement the proposed work we have used the standard dataset of 11 languages word images[1] from that we have selected Kannada, Hindi(Devnagari) and English scripts dataset which is containing 20,000 for each script in that we considered 2,000 images for each script. Though the dataset is already preprocessed so, not done any pre-processing step.

The sample input dataset are used to extract the features are

(a) Kannada (b) Hindi (c) English

Figure 1

IV. Feature Extraction

For any script identification features a play a vital role in the system. From the features only it is to discriminate the different scripts from tri-lingual scripts. For this problem the feature extraction technique should good enough to extract the dominant features from the word images. The word images are different from each other in the shape.

For extraction of features we LBP (Local Binary Pattern) texture features. This technique is compare between pixels with the neighbors, if it is greater than the neighbor then it puts 1 else 0 and generated 0 and 1 patterns is encoded.

From the encoded patterns the LBP histogram can be defined by using following equation.

$$H_i = \sum_{x,y}^{n} I\{f_1(x,y) = i\}, i = 0,...,n-1$$

After getting the histogram it has to be normalized to get a coherent description:

$$N_i = \frac{H_i}{\sum_{j=0}^{n-1} H_j}$$

The following algorithm is defined to script identification of word level.

Step 1: Start
Step 2: Input the binary word images
Step 3: Feature extraction using LBP Method
Step 4: Apply the Classifiers K-NN and SVM
Step 5: Identified script
Step 6: Stop

V. RESULTS AND DISCUSSION

The following tables are showing the average recognition accuracy for tri-lingual scripts.

Table 1: Confusion matrix for Tri-lingual Scripts using K-NN classifier with Euclidian distance Metric.

Scripts	Kannada	Hindi	English	Rec. Acc
Kannada	1975	10	15	98.75%
Hindi	34	1960	16	98.00%
English	15	17	1968	98.40%
Average Recognition Accuracy				98.38%

Table 2: Confusion matrix for Tri-lingual Scripts using SVM Classifier with Cubic kernel function

Scripts	Kannada	Hindi	English	Rec. Acc
Kannada	1971	16	13	98.55%
Hindi	26	1962	12	98.10%
English	8	15	1977	98.85%
Average Recognition Accuracy				98.50%

The below graphs shows the script wise performance of the K-NN classifier

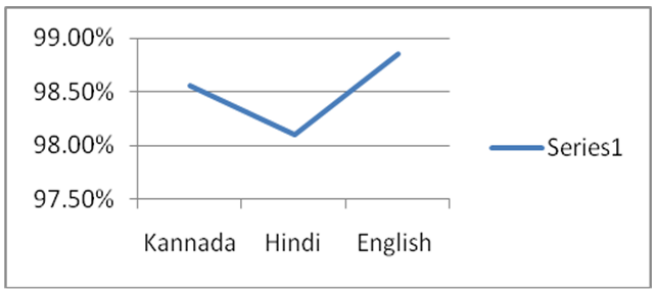

Figure 2 : K-NN classifier performance evaluation.

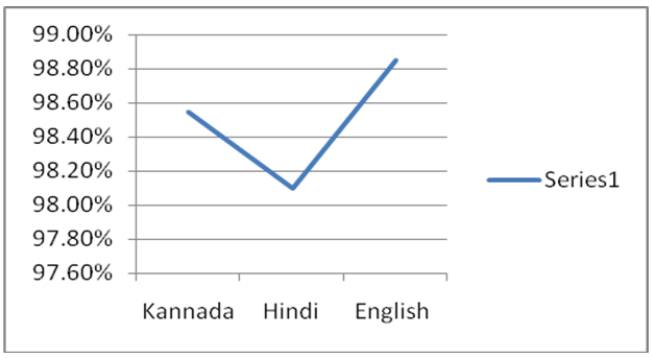

Figure 3 : SVM Classifier performance evaluation

VI. CONCLUSION

In this paper, tri-lingual script identification has been proposed, the LBP features are used to classify the scripts of Kannada, Hindi and English word-wise images. The performance of the proposed method is good to discriminate the tri-lingual scripts , and we obtained the result of 98.38% for K-NN and 98.50% for SVM. In this work we have considered only three scripts in future work we concentrate on increasing the dataset size, and addition of scripts.

VII. REFERENCES

[1]. Peeta Basa Pati and A. G. Ramakrishnan, "Word Level Multi-script Identification", Pattern Recognition Letters, 2008, Vol. 29, pp. 1218-1229.

[2]. J. Hochberg, P. Kelly, T Thomas and L Kerns, "Automatic script identification from document images using clusterbased templates," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.19, pp.176-181, 1997

[3]. Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, "Script and language identification for hand-written document images," IJDAR, vol.2, pp45-52. 1999.

[4]. S. Wood. X. Yao. K.Krishnamurthi and L.Dang "Language identification from for printed text independent of segmentation," Proc. of Int'l. Conf. on Image Processing, pp.428-431, 1995.

[5]. T.N.Tan, "Rotation invariant texture features and their use in automatic script identification," IEEE Trans.on Pattern Analysis and Machine Intelligence, vol. 20, pp.751-756, 1998.

[6]. G.S.Peake and Tan, "Script and language identification from document images," Proc. of Eighth British Mach. Vision Conf., vol.2, pp. 230-233, Sept-1997.

[7]. A.Busch ,W.W.Boles and S.Sridharan, " Texture for script identification" IEEE Trans. On Pattern Analysis and MachineIntelligence, 27(11) 1720-173,2005

[8]. Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy,"Script Identification for Indian Documents", In. Pro. of 7th IAPR workshop on Document Image Systems, (DAS), New Zealand,pp.255-267, 2006.

[9]. Peeta Basa Pati and A.G.Ramakrishnan," HVS inspired system for Script Identification in Indian Multi-Script Documents", In Proc. of 7th International Workshop on Document Analysis System, Nelson Newland,pp-380-389, Feb-13-15,2006

[10]. D Dhanya, A.G Ramakrishnan and Peeta Basa pati, "Script identification in printed bilingual documents," Sadhana, vol.27, part-1, pp. 73-82, 2002