# Word-wise South Indian Script Identification using GLCM and Radon Features

Shivanand S. Rumma
Chairman
Dept. of P.G. Studies and Research in Computer Science,
Gulbarga University, Kalaburagi.
Karnataka, India
*Email id: shivanand_sr@yahoo.co.in*

***Abstract:*** This paper presents a hybrid features for identification of south Indian scripts in word-wise and it has used three classifiers. We have used two kinds of features namely Radon and Gray Level Co-occurrence Matrix (GLCM) and combination of Radon and GLCM features. For identification purpose LDA, KNN and SVM classifiers have been employed. For the experiment proposed work considered the 6 languages scripts; Roman, Devnagari, Kannada, Telugu, Tamil and Malayalam. This proposed work considered the Word Image Dataset for 11 Languages form MILE Lab IISC in this dataset proposed work considered 6 languages with 5000 for each scripts, this makes total of 30,000 word images. We have made the total of five bi-lingual combinations of south Indian scripts. To extract features; GLCM and Radon Features are considered (4 features of GLCM, 11 features, for Radon we obtained 98.80% from KNN for the Roman and Kannada combination, for GLCM 88.20% obtained by SVM for the Roman and Kannada from SVM Classifier and from combination of Radon and GLCM we have obtained the accuracy of 98.90% for Roman and Kannada combination scripts.

***Keywords:*** *GLCM, KNN ,LDA, Radon, SVM.*

*****

## I. Introduction

The symbolic representation of the language is called Script. It is the combination symbolic representation of language each symbol has got its own characteristics. In India normally bi-lingual and tri-lingual documents may found in various states, in that it is oblivious for containing the bi-lingual scripts those are regional script and International script English( Roman). In the south India the popular scripts are Kannada, Telugu, Tamil and Malayalam with Devanagari (Hindi) and Roman (English) . If we consider the documents from south India it may contain one regional language that may be anyone from four south Indian script along with Roman script and some Government documents having Regional, National (Hindi) and International script in the document ( like Voter Id, Driving license, Post office documents etc). For such documents we need to process them for further consideration. To recognize the document, first we need to identify the script then we can feed that script to Optical Character Recognition (OCR), because OCR is a script specific. Until now the English OCR has got highest results for English documents. The English OCR has achieved the phenomenon results, whereas other scripts like Hindi, Kannada, Telugu, Tamil and Malayalam scripts still has to reach as highest as English. In this regard, there is a problem of other Script OCR to select the appropriate script OCR for processing documents containing bi-lingual scripts. So, this is the motivation for this proposed work.

## II. Literature Survey

There are significanct works has been reported in word-wise script identification. Patil et.al [11] proposed the neural network based system for English, Hindi and Kannada scripts in word-wise and they utilized the modular neural network method for classification of scripts. Dhanya et al.[2] they implemented the work on word-wise script identification using Gabor filter based technique. They have proposed a Gabor filter based technique for word-wise script identification from the bilingual documents which consisted English and Tamil scripts. Malemath et.al [3] have proposed the word wise script Identification based on Steerable Gaussian filter for printed document Images and they have used KNN classifier. Chaudhuri et al. [4] discussed an OCR system to read two Indian languages scripts: Bangla and Devnagari (Hindi). Hangarge et.al [5] proposed the word level script identification and they implemented the tool of morphological opening following by reconstruction of the images. They have considered the Kannada, Telugu and Hindi scripts. David et.al [6] have presented the comparative performance of the classifiers; SVM, KNN and GMM.

## III. Proposed Method

The proposed method utilized Radon and GLCM features, where Radon gives 8 features and GLCM gives the 4 features and when we combine these methods we obtain 12 features.

**Radon Transform**: Applying the Radon transform on an image f(x,y) for a given set of angles can be thought of as computing the projection of the image along the given angles. The resulting projection is the sum of the intensities of the pixels in each direction, i.e. a line integral. The result is a new image R(ρ,θ).

The mathematical form is :

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

Radon Transform is shown as:

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

Where $\delta$ (.) is the Dirac delta function

**Gray Level Co-occurrence Matrix (GLCM) :** The GLCM derives the Statistical properties of the image. The following are the properties of the image:

(a) **Contrast :** it gives the measurement of the intensity contrast between a pixel and neighboring pixel on the full image. Mathematically we can present as:

$$\sum_{i,j} |i,j|^2\, p(i,j) \qquad (3)$$

(b) **Correlation :** Correlation of the pixel with the neighbor pixel in the image.

$$\sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j} \qquad (4)$$

(c) **Energy:** sum of squared elements in the Image.

$$\sum_{i,j} p(i,j)^2 \qquad (5)$$

(d) **Homogeneity :** This gives the value that calculate the closeness of the distribution of elements for the GLCM to the GLCM diagonal

$$\sum_{i,j} \frac{p(i,j)}{1+|i-j|} \qquad (6)$$

Algorithm for Radon and GLCM feature extraction

```
Step 1: Start
Step 2: Preprocessed Binary Input Image
Step 3: Compute the Contrast
Step 4: Compute the Corrleation
Step 5: Compute Energy
Step 6: Compute Homogeneity
Step 7: Generate 4 features from above
        steps 2 to 5
Step 8 :Calculate the Radon Transformation
        for the Input image.
Step 9: From step 7 9 features are generated.
Step 10: Combine and GLCM and Radon
        Features, total of 13 feature vector is
        created.
Step 11:To identify the script, feed the
        features to LDA, K-NN, and SVM
        Classifier with 2-fold Cross
        validation.
Step 12:Stop
```

## IV. Results and Discussions

For the proposed experiments we have Considered standard dataset of Word Image Dataset for 11 Languages form MILE Lab IISC Bangalore, which is freely available dataset. From the dataset we have considered only 6 languages namely; Roman (English), Devenagari (Hindi), Kannada, Telugu, Tamil and Malayalam. The following are the input images

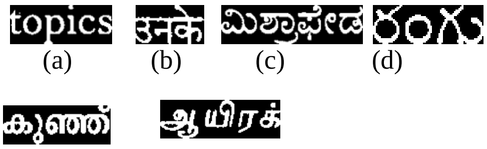

Fig 1: Input images. a) Roman b) Devnagari
c) Kannada d) Telugu e) Tamil
f) Malayalam

Following table 1-3 shows the accuracy of the popular south Indian scripts.

Table 1: Average Recognition Accuracy of LDA , KNN and SVM Classifier with 2-fold Cross Validation for Bi-lingual South Indian Words document image by GLCM Features.

| **GLCM** | | | |
|---|---|---|---|
| Scripts/Classifier | LDA | KNN | SVM |
| R-H | 67.90% | 82.20% | 83.50% |
| R-K | 70.70% | 87.80% | **88.20%** |
| R-Te | 61.50% | 74.30% | 75.06% |
| R-Ta | 72.00% | 82.50% | 85.50% |
| R-M | 72.50% | 82.80% | 83.30% |

From the above table 1, it is observed that the Romanwith Kannada Combination is obtained 88.20% accuracy from SVM Classifier.

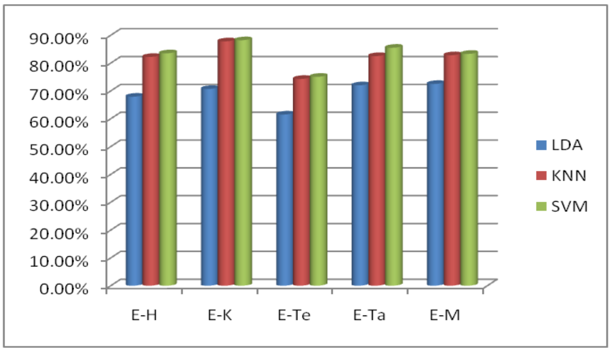

Fig 1: Results of GLCM Features with LDA, KNN and SVM Classifiers.

Table 2: Average Recognition Accuracy of LDA , KNN and SVM Classifier with 2-fold Cross Validation for Bi-lingual South Indian Words document image by RADON Features.

| RADON | | | |
|---|---|---|---|
| Scripts/Classifier | LDA | KNN | SVM |
| R-H | 87.70% | 94.90% | 90.80% |
| R-K | 95.80% | **98.80%** | 98.30% |
| R-Te | 94.50% | 98.50% | 97.80% |
| R-Ta | 87.20% | 94.70% | 92.10% |
| R-M | 85.40% | 9.32% | 90.40% |

The above presented table clearly shows the highest result of 98.80% accuracy for Romanand Kannada features by KNN Classifier.

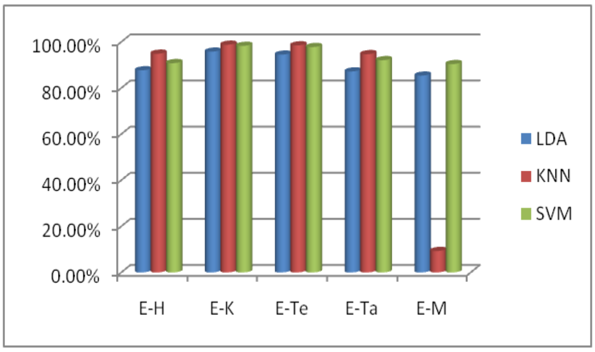

Fig 2: Results of RADON Features with LDA, KNN and SVM Classifiers.

Table 3: Average Recognition Accuracy of LDA , KNN and SVM Classifier with 2-fold Cross Validation for Bi-lingual South Indian Words document image by combining the GLCM+RADON Features

| GLCM+RADON | | | |
|---|---|---|---|
| Scripts/Classifier | LDA | KNN | SVM |
| R-H | 87.50% | 97.40% | 96.10% |
| R-K | 95.10% | **98.90%** | 98.80% |
| R-Te | 93.60% | 98.70% | 98.40% |
| R-Ta | 86.50% | 94.50% | 93.10% |
| R-M | 84.40% | 94.20% | 94.50% |

The above table shows the 98.90% highest accuracy for the Roman and Kannada scripts with KNN classifier.

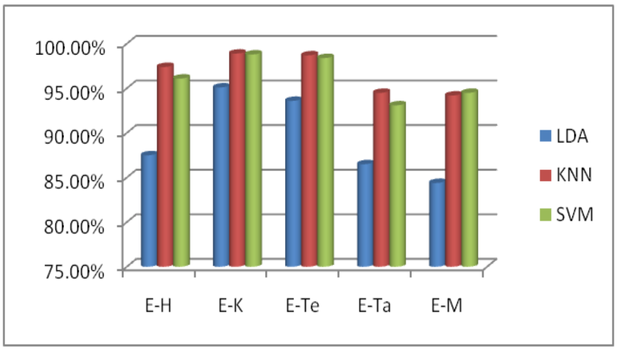

Fig 3: Results of RADON + GLCM Combined Features with LDA, KNN and SVM Classifiers.

By observing the above tables and figures it is shown that the combination of the GLCM and RADON features are the potential features for Kannada and Romanscripts, for other combination one need to add more potential features to reach the as highest as 100% accuracy.

## V. Conclusion

This paper presented the performance analysis of Radon and GLCM features along with LDA, KNN and SVM Classifiers. The results are obtained by using combining two features and it has given positive results the 13 features were used for the recognizing the scripts belonging to six different scripts. The proposed work is obtained the optimum result of 98.90% with the combination of Roman and Kannada with the features combining GLCM and RADON. In future work we extend the no. of scripts and increase the dataset and obtain the results with minimum features and highest accuracy.

## References

[1] S. B. Patil and N. V. Subbareddy, Neural network based system for script identification in Indian documents,Sadhana, vol. 27, pp. 83-97, 2002.

[2] D. Dhanya, A. G. Ramakrishna, and P. B.Pati, "Script identification in printed bilingual documents," Sadhana, vol. 27, pp. 73-82, 2002.

[3] V .S. Malemath, A. H. Kulkarni and H. Mallikarjun, Word-wise Script Identification in Document Images based on Steerable Gaussian Filtering Technique ,International Journal of Advanced Research in computer and communication and Engineering" ,vol 3,no.6,Jun 2014

[4] B.B.Chaudhuri and U.Pal," An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi),Proc. of 4th ICDAR, Uhn. 18-20 August, 1997.

[5] M. Hangarge and B.V .Dhandra, "Morphological Reconstruction for Word level script identification", International Journal of Computer Science and Security, vol.1, no.1, pp 41-51.

[6] David Doermann and Huanfeng Ma, "word level script identification for scanned images", Language and Media Processing Laboratory Institute for Advanced Computer Studies University of Maryland, College Park, MD 20742, USA.

[7] Rafael Gonzalez and R Woods, IInd Ed., Digital Image Processing, Pearson Education, 2004.

[8] B V Dhandra et. al, Word-level Script Identification in Bilingual Documents through discriminating features", In the Proc. Of International Conference on Signal Processing Communications and Networking (ICSCN2007) Chennai held during 22-24 Feb.2007.

[9] G.Mukarambi et al. Script Identification from Camera Based Tri-lingual Document,ICSSS, IEEE, pp.214-217. May.2017

[10] P.Nagabhushan, S.A.Angadi, and B.S.Anami, An Intelligent Pin code Script Identification Methodology Based on Texture Analysis using Modified Invariant Moments, In Proceedings of International Conference on Cognition and Recognition, pp. 615-623, 2005

[11] 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.

[12] https:// en. wikipedia.org /wiki/Linear_discriminant_analysis