International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Multi-Feature based Handwritten Script Identification at word level

Suryakanth Baburao Ummapure, G. G. Rajput

Abstract: ***SIFT and LBP are two popular techniques used for obtaining "feature description" of the object. SIFT identifies key points that are locations with distinct image information and robust to scaling and rotation whereas, LBP transforms an image into an array of integer labels describing small scale appearance of the image. In this paper, we present an efficient method wherein "feature description" of handwritten document images at word level are computed using SIFT and LBP. Identification of script type is done using KNN and SVM classifiers. Experimental results show that the performance of SVM is better over KNN. Further, the proposed method is compared with other methods in the literature to demonstrate the efficacy of the proposed method***.

Keywords : Script identification, word level, SIFTS, LBP, KNN, and SVM.

I. INTRODUCTION

Optical character recognition (OCR) system is used in document image analysis for character recognition. The input to OCR is text document of a specific script. Script type identification plays a significant role of identifying the script type of the text, from mono/multi-lingual text document, and feeding the text to appropriate OCR for character recognition. Many methods have been proposed in the literature for script type identification from a multiscript/monoscript document images [1]. Significant work has been carried out for printed documents compared to handwritten documents [2]. Further, in a country like India where more than one script is in use for communication, script identification is a must to facilitate OCR system of respective language. Script identification, is generally carried out at one of the four levels, namely, page level, block level, line level and at word level. A typical script identification process at word level is shown in Fig.1 . Pre-processing involves noise removal, binarization, removal of irrelevant details, and extraction of the word from the document image. The word image is then processed to compute features. In the training phase, features are computed for all the word images and these labeled features are stored as knowledge base.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Suryakanth Baburao Ummapure*, Department of Computer Science Gulbarga University, Kalaburagi, 585106, Karnataka, India Email: ummapure@gmail.com
G G Rajput, Department of Computer Science Akkamahadevi Women's University, Vijayapura 586106, Karnataka, India, Email: ggrajput@yahoo.co.in

During test phase, features computed from test word image are compared with the stored features of the knowledge base using pre-defined criteria and the outcome is the label of the selected stored feature indicating the script type of the test image.

Literature survey reveals that most of the script type identification is based on one type of features extracted from the word image [3]. Very few works are found employing more than one feature [4]. In this paper, script identification at word level, extracted from handwritten document images, is presented by combining features obtained from SIFT and LBP techniques. Experimental results obtained are encouraging and a very few approaches of this kind are found in the literature.

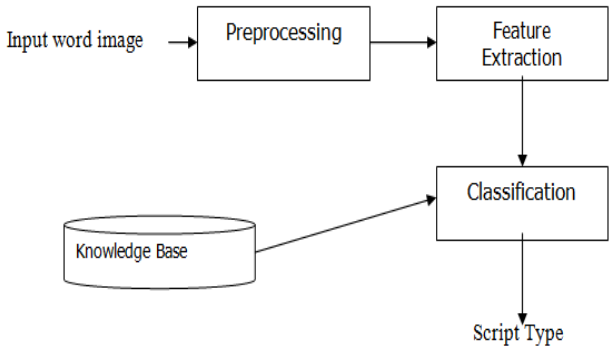

Fig 1: Block diagram of Script Identification

II. LITERATURE REVIEW

A brief overview of hand written script identification using multiple features is presented below.

G.G. Rajput et al., in [5] reported hand written script identification using multiple features based approach at line level. Features are extracted using Gabor filter, DCT, wavelets of Daubechies family script type classification has been performed using KNN and SVM.SK. Md. Obaidullah et al., in [6] reported script identification methodology from Indian hand written documents Bangla, Roman, Malayalam, Urdu, Oriya and Devanagari script used to extract mathematical, structural, script dependent features, circularity, fractal dimension features were also computed using MLP classifier.

HOG and Gabor feature based Arabic hand written script identification proposed by Mohamed Elleuch et al., in [7]. Experiments are performed on Arabic hand written database namely IFN / ENIT. SVM classifier used to recognize the script. Alia Karim Abdul Hassan et al.,[8] reported different features extraction methods i.e., DWT, profile based and direction features. Experiments performed on IESK- arDB database and classification performed using KNN.

Namboodiri et al., proposed online script recognition in [9] for Arabic Cyrillic, Devanagari, Han, Hebrew and Roman scripts. Spatial and temporal features of strokes were extracted. Namely HID, ASL, Shirorekha strength, Shirorekha confidence (SC), stroke density (SD), Aspect ratio (AR) average horizontal stroke direction (AHSD) and average vertical stroke direction (AVSD) and have used SVM classifier to classify different hand written scripts.

Effectiveness features extracted using Gobar and DCT is proposed by Peeta Basa Pati et al., in [10] to identify bi-script, tri-script and multi scripts. Nearest neighbor, linear discriminant and SVM classifier used to estimate the effectiveness. NN and SVM classifiers have shown better recognition with the combination of Gabor features. In [11] texture as a tool has been proposed by Hangarge et al., to identify hand written script based on well defined visual texture. KNN classifier used to perform experiment at different K values on the Indian scripts i.e., English, Devanagari and Urdu. Morphological features were used to extract 13 spatial spread features.

Pawan Kumar Singh et al., in [12] have reported word level script identification from Malayalam, Oriya, Telugu, Tamil and Roman hand written documents. A total of 92 feature vector created out of which 80 features are extracted using DCT and 12 feature are selected from moment invariant. MLP and SVM classifiers used to classify 1000 words database. MLP classifiers shown better performance over SVM. Shivanand S. Rumma reported multi feature based south Indian script identification from printed document in [13] and used standard database[MILE Lab IISC Bangalore] to test performance of the proposed method. GLCM and Radon based features have been extracted for Roman, Devanagari, Kannada, Telugu, Tamil and Malayalam, the scripts are classified using LDA, KNN and SVM classifiers. Proposed method performance is better for Roman and Kannada scripts. D.S. Guru et al., reported a survey on offline hand written script recognition in [14] which includes various algorithms proposed to extract features to identify hand written scripts with combined features to obtain promising results.

III. DATA COLLECTION AND PREPROCESSING

Script identification for handwritten documents written in Kannada, Hindi and English script is considered in this paper. To the best of our knowledge, standard database for handwritten script image at word level for Kannada script is not available. Hence, we have created a large collection of word images of Kannada script and for the scripts written in Hindi and English. For this, hand written documents are created by persons from different age groups and professions. A total of 300 documents images were collected. The details of the documents is shown in the Table 1.

Table- I : Handwritten documents details

Sl. No.	Script Type	Number of document pages created	Sources
1.	Kannada	100	Persons of different age groups and professions with formal communication practice in English, Kannada and Hindi languages
2.	Devanagari	100	
3.	English	100	

The collected documents are scanned using HP flatbed scanner. Preprocessing of the document images and extraction of the words from the document images is performed by applying the methods proposed in our earlier work [15][16]. In order to extract the words from the segmented text line we incorporate a bounding box expansion technique to extract a meaningful text word present in the segmented text line.

IV. FEATURE EXTRACTION AND CLASSIFICATION

Features are extracted using two approaches namely, SIFT [17] and LBP [19]. A brief description of the techniques is presented below.

Scale Invariant Feature Extraction

In SIFT; initially detect scale space by finding scale and locations using DOG operation subject to intersecting points. Eliminate the pixels having low contrast. Using local image, compute magnitude gradient and orientation angle for the smooth images to define highest peak point. Assign the key points based on orientation by calculating difference of pixel from the histograms with respect to dominant orientations. Finally, key point descriptor with local gradient data is created. The key point descriptors obtained using the images are concatenated to form a feature vector of 128 elements [17].

Local Binary Patterns

The word image is divided into 3x3 cells and LBP operator is applied to compute the compact representation of the image as binary code. Histogram is generated for each cell and the histogram is normalized. The resulting values for each cell are concatenated to obtain a vector of 59 elements [19].

Finally, LBP features and SIFT features are concatenated to form a single feature vector. The algorithm for feature extraction is presented below.

Algorithm :

Input : preprocessed word image

Output : Feature vector computed from SIFT and LBP operator

1. Input the pre-processed image.
2. Apply SIFT method to compute feature vector of size 128 elements [17].
3. Apply LBP operator and extract features vector of size 59 elements [19].
4. Combine the feature obtained in steps 2 and 3 to form a feature vector of size 187.
5. Repeat the steps 1-4 for all the images of training set to obtain feature vectors for word images of training set. Label each feature vector with the label of the script type. This collection of labeled feature vectors forms the knowledge base.
6. During test phase, perform steps 1-4 for the input text word image to obtain feature vector.
7. Input the features of the test image obtained in step 6 and the knowledge base to KNN classifier for script identification. K-NN outputs the label of script type of the test image.
8. In case of script identification using SVM, during the training phase the feature vectors obtained in step 5 for training images are input to SVM to generate support vectors.

During testing phase, the features extracted from test image (step 6) are input to SVM for script type identification.

A feature matrix of size 70 x 187 for bi-script and 100 x 187 for tri-script is used for performing experiments.

V. EXPERIMENTAL RESULTS

The performance of the proposed algorithm is tested on the scripts written in Kannada, English and Devanagari. For performing experiments, 630 text words at bi-script level and 510 text words at tri-script level are considered. The results obtained for KNN and SVM classifiers in terms of recognition accuracy and confusion matrix are shown in the tables 1 through 6. The results are presented in these tables for word images consisting of two, three and more than three character words, respectively. SVM yielded better results compared to KNN classifier. The proposed method is compared with the multi-feature based feature extraction methods in literature. The comparative study in terms of recognition accuracy is presented in Table 7. The results are comparable with other methods and that the proposed method demonstrates the efficacy of LBP and SIFT over other methods.

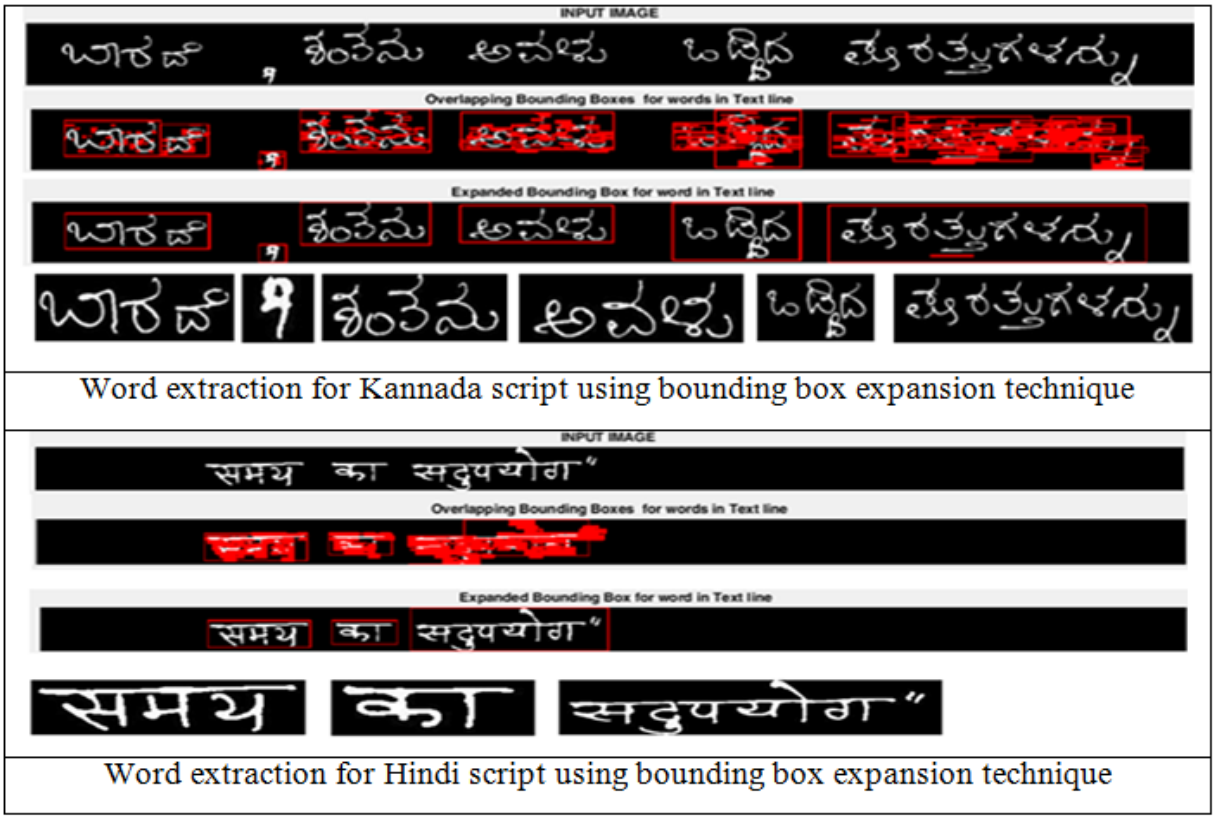

Fig 2: Sample image for Word extraction from Kannada and Hindi Script

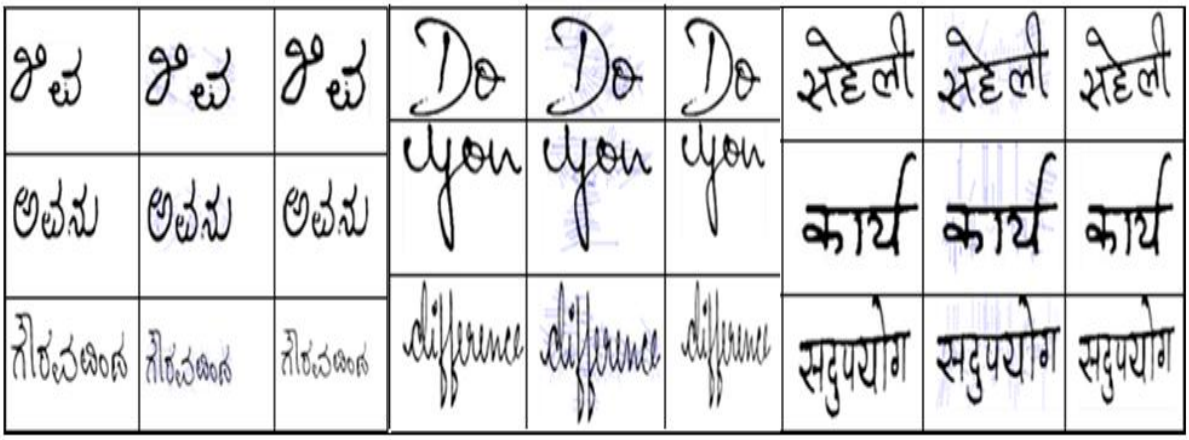

Fig 3: Sample images for Key point localization using SIFT method

Table- II: Confusion matrix and recognition accuracy for Bi-Script using KNN and SVM

Sl.No	Script Type	Two Character Word		KNN	Two Character Word		SVM	Three Character Word		KNN	Three Character Word		SVM
1	Kannada	65	5	94.3	70	0	100	68	2	97.9	69	1	99.3
	English	3	67		0	70		1	69		0	70	
2	Kannda	55	15	78.5	66	4	90.7	55	15	86.42	67	3	96.4
	Hindi	15	55		9	61		4	66		2	68	
3	English	70	0	98.6	70	0	100	69	1	99.3	70	0	100
	Hindi	2	68		0	70		0	70		0	70	
Over All reciginition accuracy in %				90.47			96.90			94.54			98.57

Sl.No	Script type	More than Three character word		KNN	More than Three character word		SVM
1	Kannada	88	12	91	93	7	95.5
	English	6	94		2	98	
2	Kannada	88	12	85.5	94	6	95
	Hindi	17	83		4	96	
3	English	97	3	92	100	0	100
	Hindi	13	87		0	100	
	Over all Recognition accuracy in %			89.50			96.83

Table- III: Over all recognition accuracy for Bi-Script

Words	KNN	SVM
Two Character Word	90.47	96.9
Three Character Word	94.54	98.57
More than Three Character Word	89.5	96.83
Recognition accuracy In %	91.50	97.43

Table-IV : Confusion matrix and recognition accuracy for Tri-Script using KNN and SVM

Script Type	Two Character Word			KNN	Two Character Word			SVM
English	52	5	13	81.52	64	0	6	92.37
Hindi	5	65	0		0	70	0	
Kannada	16	0	54		10	0	60	

Script Type	Three Character Word			KNN	Three Character Word			SVM
English	46	1	23	80.5	66	1	3	96.7
Hindi	3	67	0		0	70	0	
Kannada	14	0	56		3	0	67	

Script Type	More than Three Character Word			KNN	More than Three Character Word			SVM
English	51	40	9	74	93	2	5	94.3
Hindi	10	87	3		2	97	1	
Kannada	7	9	84		7	0	93	

Table- V: Over all recognition accuary for Tri-Script

Words	KNN	SVM
Two Character Word	81.52	92.37
Three Character Word	80.5	96.7
More than Three Character Word	74	94.3
Over all Recognition accuracy in %	78.67	94.46

Table- VI: Over Accuracy for Bi-script and Tri-script

Sl. No	Script Type	KNN	SVM
1	Bi-Script	91.50	97.43
2	Tri-Script	78.67	94.46
Over all Recognition Accuracy in %		85.08	95.94

Table- VII: Comparative Study

Authors	Features	Classifier	Result
SK. Md. Obaidullah et al.,[6]	Mathematical, Structural, Script Dependent, Circularity, Fractal Dimensions	MLP	92.8%
Namboodiri et al.,[9]	Spatial and Temporal Features of Strokes	SVM	95%
Peeta Baba Pati et al.,[10]	Gabor and DCT	NN&SVM	89%
Hangarge et al.,[11]	Vertical, Horizontal, Right and Left Diagonal Stroke Density	KNN	Bi-script 99.2% Tri-Script 88.6%
Pawan Kumar Singh et al.,[12]	DCT and Moment Invariant Features	MLP	93.56%
Proposed method	**SIFT and LBP**	**SVM**	**95.94%**

VI. CONCLUSION

In this paper an efficient approach for hand written script identification at word level is proposed based upon the features obtained using SIFT and LBP methods. The experiments are performed at two levels, namely, bi-script and tri-script, using KNN and SVM classifiers. Performance of SVM classifier is better over KNN classifier. The proposed method can be extended to other scripts at both word level and line level.

REFERENCES

1. Sahare, P., & Dhok, S. B. (2017). Script identification algorithms: a survey. International Journal of Multimedia Information Retrieval, 6(3), 211–232. doi:10.1007/s13735-017-0130-2.
2. Pawan Kumar Singh, Ram Sarkar, Mita Nasipuri,Offline Script Identification from multilingual Indic-script documents: A state-of-the-art ,Computer Science Review, Volumes 15–16, February – May 2015, Pages 1-28.
3. K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo, T. Yibulayin , " Script Identification of Multi-Script Documents: a Survey " , IEEE Access DOI 10.1109/ACCESS.2017.2689159 volume 5 2017, pp. 6546-6559.
4. D S Guru , M Ravikumar and B S Harish , "A Review on Offline Handwritten Script Identification", International Journal of Computer Applications (0975 – 8878) on National Conference on Advanced Computing and Communications - NCACC, April 2012, pp.13-16.
5. G G Rajput and Anita H B Handwritten Script Recognition at Line Level – A Multiple Feature Based Approach, International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 4, October 2013 ISSN: 2277-3754 pp 90-95.
6. Obaidullah, Das & Roy, System for Handwritten Script Identification from Indian Document, Journal of Pattern Recognition Research 8 (2013) 1-12, pp 1-12.
7. Mohamed Elleuch, Ansar Hani, and Monji Kherallah, Arabic Handwritten Script Recognition System Based on HOG and Gabor Features, The International Arab Journal of Information Technology, Vol. 14, No. 4A, Special Issue 2017, pp 639-646.
8. Alia Karim Abdul Hassan and Mohammed Alawi, Proposed Multi Feature Extraction Method for Off-line Arabic Handwriting Word Recognition, Al-Mansour Journal/ Issue (30) 2018, pp-17-31.
9. Anoop M Namboodiri and Anil K Jain., 2002, On- line script recognition. 16th International Conference on Pattern Recognition, pp 736-739 .
10. Peeta basu patil and A G Ramakrishnan., 2008. Word level multi-script identification. Pattern recognition letters, Vol 29, pp 1218 – 1229.
11. Mallikarjun hangare and B V Dhandra., 2010. Offline handwritten script identification in document images. Intenational journal of computer applications. Vol 4, pp 6 – 10.
12. Pawan Kumar Singh, Aparajita Khan, Ram Sarkar, Mita Nasipuri, A Texture based approach to Word-level Script Identification from Multi-script Handwritten Documents ,2014 Sixth International Conference on Computational Intelligence and Communication Networks, pp-228-232.
13. Shivanand S. Rumma, Word-wise South Indian Script Identification using GLCM and Radon Features, International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 2 476 – 478 IJFRCSCE | February 2018, Available @ http://www.ijfrcsce.org.
14. D S Guru, M Ravikumar and B S Harish, A Review on Offline Handwritten Script Identification, International Journal of Computer Applications (0975 – 8878) on National Conference on Advanced Computing and Communications - NCACC, April 2012,pp-13-16.
15. G. G. Rajput , Suryakant B. Ummapure and Preethi N. Patil, " Text Line Extraction from Handwritten Document images using Histogram and Connected Component Analysis," International Journal of Computer Applications (0975 - 8887) National conference on Digital Image and Signal Processing , D1SP 2015. Pp 11-17.
16. G. G. Rajput , Suryakant B. Ummapure and Panditkumar Patil, "Separation of Touching or Overlapping Lines from Handwritten Document images using Histogram and Connected Component Analysis," International Journal of Computer Applications (0975-8887)National Conference on Digital Image and Signal Processing 2016.
17. G. G. Rajput and Suryakant Baburao Ummapure "Script Identification from Handwritten Documents using SIFT Method" IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-20 17) 978-1-5386-0814-2/17/$31.00 ©2017 IEEE, pp 520-526.
18. G Rajput and Suryakant Baburao Ummapure " Line-wise Script identification from handwritten document images using SIFT method" Proceedings of the International Conference on Intelligent Computing Systems (ICICS 2017 – Dec 15th -16th 2017) organized by Sona College of Technology, Salem, Tamilnadu, India Elsevier's SSRN eLibrary – Journal of Information Systems &eBusiness Network -ISSN: 1556-5068 pp117-125.
19. G. G. Rajput, Suryakanth Baburao Ummapure, Script Identification from Handwritten Document Images Using LBP features, ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 9|| September – 2018 ||International Journal of Computational Engineering Research (IJCER),pp-13-21.

AUTHORS PROFILE

Suryakanth Baburao Ummapure is pursuing Ph.D in Computer Science from Gulbarga University Kalaburagi and working as Assistant Professor at Government First Grade College Shahapur Dist. Yadgiri. His area of Interest is Image processing. He has presented many papers in National and International Conferences, published research papers in National/ International Journals/ Conference Proceedings.

Dr. G. G. Rajput working as Professor with Karnataka State Akkamahadevi Women's University, Vijayapura. His area of research work includes Image processing and Pattern Recognition, Document Image Processing, Biometrics, Medical Image Processing and Data Mining. He has published many research papers in National and International Journals.