ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Word-wise Script Identification in Document Images based on Steerable Gaussian Filtering Technique

V. S. Malemath[1], A. H. Kulkarni[2], H. Mallikarjun[3]

Department of CSE, KLE DR. M.S. Sheshgiri College of Engineering & Technology, Belgaum, Karnataka, India[1]

Department of CSE, KLS's Gogte Institute Technology, Belgaum, Karnataka, India[2]

Department of CSE, KASCC, Bidar, Karnataka, India[3]

Abstract: In this paper, a study on word wise script identification based on Steerable Gaussian filter for printed document images is carried out. The system is developed and tested for 3000 document image data set representing English, Hindi, Kannada, Tamil and Urdu script word images of 600 each. The system developed includes a feature extractor which is based on Steerable Gaussian filter technique and for classification K-nearest neighbor classifier and linear discriminate classification techniques are used. The feature extractor consists of application of steerable Gaussian filter at different orientations 0, 45, 90, 135, 30, 65, 155 and the associated standard deviation of the local orientation is used as the feature set thus contributing only seven features. The two classifications techniques were used for analysis of the new word-wise segmented documents. Classification accuracy averaged 97% across the five scripts .The method shows robustness with respect to noise, the presence of headlines, font sizes and styles.

Key words: Document image processing, steerable filter, script, Identification, OCR

I. INTRODUCTION

The OCR technology is of special significance with recent trends for paperless office and in a multi-lingual country like India the problem becomes even more dominant. Although a large number of OCR techniques have been developed over the years, almost all-existing works on OCR assume that the script and language of the document is known beforehand. Thus, individual OCR tools have been developed to deal with only one specific language. That is, an OCR developed for English will work satisfactorily for the English document with desired accuracy, where as this system not at all work satisfactorily for the documents with other local scripts like Hindi, Kannada etc. In regard to this, there is a need to develop pre-OCR script and language identification system to enable to select the appropriate OCR system for processing the document containing different scripts and languages.

In a multi-script, multi-lingual country like India (India has 18 regional languages derived from 12 different scripts [7]), now there is a growing demand for automatic processing of the document in every state in India including Karnataka. Under the three-language formulae [7], the documents in a Karnataka state may be printed in its respective official regional language Kannada, the national language Hindi, and also in English. Further the documents produced in north Karnataka state are influenced by Urdu scrip in addition to Kannada, Hindi and English, since this part is ruled by Nizam and with Tamil at southern region. This has motivated us to propose a method for automatic script or language identification from document images containing any of these proposed scripts.

In the literature it is found that most of the work carried on the script identification is in three fold i.e. in the first fold the entire document page is in a mono script and used in identification of the script. Some piece works have also been reported on block by block of the size of 128X128 or 256X256 sized images of the mono script.

The second type of work reported is on the line level. In this type the one text line of the document is extracted and its script is identified. Finally in the third type the work reported is on the word level. It is by far presumed as the most difficult type of script identification amongst the three ways.

Most of the Indian script identification task at word level for different configurations like doublets and triplets is carried out by U.Pal et al.[3-6]. Most of their algorithms are based on topological and structural features like number of loops, headline feature, stokes, water reservoirs and projection profiles etc. Padma et al[9-10]. have proposed the script identification improved schemes based on visual discriminating features like directional strokes, equal and unequal sized blocks and variable sized characters etc. For classification, they have considered Tamil, Telugu, Devanagari, English and Kannada.

Patil et.al[11] have devised a neural network based system for word-wise script identification of English, Hindi and Kannada scripts and they have used modular neural network technique for the classification of the proposed scripts. Dhanya et al.[12] have proposed a Gabor filter based technique for word-wise script identification from bilingual documents which consisted English and Tamil

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

scripts. Kumar et al.[28] have proposed a character-wise script identification scheme where neural network is used for the purpose. Kunte et al[29]. have proposed a script identification method to identify Kannada and English scripts from a scanned bilingual document, based on Gabor features. Radial Basis Function (RBF) Neural classifiers have been used for the classification of the scripts based on the Gabor features. Nagabhushan et al.[13] have devised an intelligent technique for PINCODE script identification using a least square distance measure to the statistical average of texture features. The texture features are defined using normalized modified invariant moments.

The methodology is tested for various machine printed scripts namely Roman, Devanagari, Kannada, Tamil and Malayalam in various fonts and sizes. Handwritten PINCODES in Roman script are also considered. Peeta Basapati et al.[14,15] have used global approach based on Gabor filter bank having three different radial frequencies and six different angles of orientation with a radial frequency bandwidth of 1 octave and an angular bandwidth of $30^{o.}$ They obtained a combination of 18 odd and 18 even filters with three radial frequencies and six degrees. The size of each filter mask used for experimentation was 13x13.

They have used a 36-dimensional feature vector of the total energy in each of the filtered images. The Linear discriminate analysis(LDA) and nearest neighbour (NN) classifiers are used to classify the word images of five different scripts namely Roman, Devnagari, Kannada, Tamil and Oriya in bi-script, tri-script and five-script levels. Vipin Gupta et al.[37] have discussed a method for script identification in trilingual documents containing Kannada, Devanagari and Roman based on character classification using language inherent features like cavities, corner points, end point connectivity and line detection.

In this paper, word-wise identification of script is carried out based on steerable Gaussian filtering technique at different orientation and is described what follows. The paper is organized as follows. The proposed method is described in section II. The steerable Gaussian filter is described in section III. Algorithm is presented in section IV. Experimental results are presented in V and the paper is concluded in VI.

II. PROPOSED METHODOLOGY

Feature extraction is the integral part of the any recognition system. The aim of feature extraction is to identify patterns by means of minimum number of features that are effective in discriminating pattern classes. The new algorithm is inspired by a simple observation that every script defines a finite set of text patterns, each having a distinct visual appearance [4] and hence every language could be identified based on its discriminating features.

It can be observed that the presence of vertical, horizontal, right diagonal, left diagonal strokes and holes in the characters of the five scripts are more distinct. Most of the Hindi (Devanagari) language characters have horizontal (sirorekha in Devanagari [7]) and vertical stokes like structure [9]. It has been found that a distinct property of the English characters is the existence of the vertical strokes like structure [7] and less number of horizontal strokes comparative to other scripts.

It could be seen that most of the Kannada characters have horizontal stokes like structures [9] and also strokes in right and left diagonal directions, where as Urdu characters contains more strokes in right diagonal and horizontal directions. The Urdu characters have less number of holes compared to English, Kannada and Hindi. Further, in the Tamil script there is a strong dominance of horizontal strokes. These discriminating features present in the characters of each script which are extracted by directional energy function which is based on Steerable Gaussian filter. In order to capture the energy at different levels as mentioned above the steerable Gaussian is used at different orientations viz. i.e. 0, 45, 90, 135, 30, 65, 155 and is described in following section.

III. STEERABLE GAUSSIAN FILTER

The proposed method is based on the extension of basic Gabor filtering technique known as Steerable Gaussian filter. The oriented filters are useful in many earlier computer vision and image processing tasks.

It's often required in some applications to apply the same filter rotated to different angles under adaptive control or wishes to calculate the filter response at various orientations.

In script the similar situations as discussed earlier are identified hence it becomes the dominant feature in the script identification problem and hence it is primarily considered for implemented. The steerable filters may be designed in Quadrature pairs to allow the adaptive control over phase and well as orientation. The concept of *steerability* was first proposed by Freeman et. al [28] and was further discussed by others in [29, 30]. A function $f(x, y)$: $R^2 \rightarrow$ C is *steerable* with respect to rotation if:

$$f^{\theta}(x,y) = \sum_{j=1}^{M} K_j(\theta)\varphi_j(x,y) --[3.1]$$

Here $f^{\theta}(x,y)$ is the rotated version (by an angle θ) of $f(x, y)$. $\varphi_j(x,y)$ (for J= 1….. M) are the base functions which are independent of the rotation angle θ. $K_j(\theta)$ (for J= 1….. M) are called the steering functions of f associated with the base functions $\varphi_j(x,y)$ and depend solely on θ. It is well known that convolution is a linear operation.

Therefore, if a filter is steerable with respect to rotation, the filter output of a rotated version of this filter can be obtained by linearly combining the filter outputs of its associated basis functions, or specifically

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

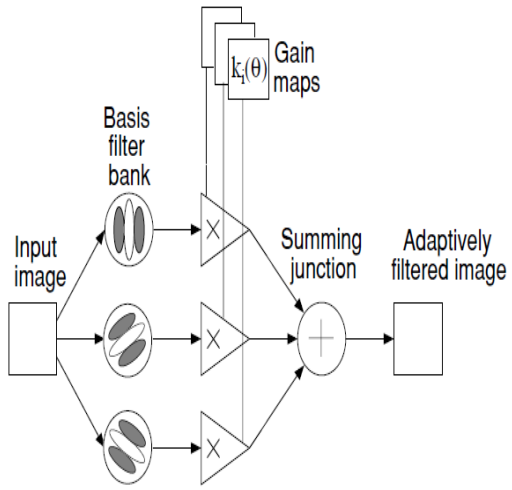

Fig. 1 Architecture of Steerable filters

The orientation strength in a particular direction by a squared output of Quadrature pair of band pass filters steered to the angle θ This spectral power is called as oriented energy and usually denoted as E(θ).

In the implementation of the algorithm the analysis of local orientation is used at the different angle i.e. 0, 45, 90, 135, 30, 65, 155. And the associated standard deviation of the local orientation is used as the feature thus contributing only seven features. For the classification the two classifiers viz. Linear discriminate analysis and the Kth Nearest Neighbor are used for the comparison.

The algorithm for the same is given as follows

IV. ALGORITHM

- Pre-process the input document image i.e. conversion to gray and binary using Otsu's method.
- Assign the default orientation at angles 0, 45, 90, 135, 30, 65, 155.
- Carry out the filtering using the designed Steerable Gauss filter at each of the orientation mentioned and estimate the spectral power / oriented energy.
- Compute the standard deviation of this oriented energy function for each oriented angle and utilize them as a feature set for training and classification
- Classify five scripts based on linear discriminate analysis (LDA) and using K-nearest neighbor classifier (KNN) classifier.

V. EXPERIMENTAL RESULTS

In order to carry out the experimentation on the document images a large set of images in the specific scripts were essential. Since there are no readymade standard databases available we have created our own databases for Indian scripts and languages from various books, magazines, weeklies, journals and newspapers. Further, online newspapers and documents were downloaded, printed and then scanned. Some document images are also downloaded from digital libraries. Documents collected have lot of variability in terms of font style, font size, and the age. The word images that are used totally 600 images per script out of which 100 images were used in training and the rest 500 images of the word were used. Thus total word images used in the experimentation are 3000.The method described showed and encouraging results in terms accuracy. Table 1 shows the script identification results for all 5 scripts. The average identification is found to be 99.125% with LDA classifier. The confusion tables of the Bi-script classification are shown for all the scripts and are represented in the Tables 2 to 5.

TABLE 1: BI-SCRIPT CLASSIFICATION RESULTS WITH LDA AND 10 FOLD CROSS VALIDATION

Script	Kannada	Hindi	Urdu	Tamil	Average %
English	98.0	99.5	100.0	99.0	99.125

TABLE 2: CONFUSION TABLE OF ENGLISH AND KANNADA SCRIPT CLASSIFICATION

Scripts	English	Kannada
English	492	8
Kannada	8	492

TABLE 3: CONFUSION TABLE OF ENGLISH AND HINDI SCRIPT CLASSIFICATION

Scripts	English	Hindi
English	497	3
Hindi	4	496

TABLE 4: CONFUSION TABLE OF ENGLISH AND URDU SCRIPT CLASSIFICATION

Scripts	English	Urdu
English	500	0
Urdu	0	500

TABLE 5: CONFUSION TABLE OF ENGLISH AND TAMIL SCRIPT CLASSIFICATION

Scripts	English	Tamil
English	495	5
Tamil	5	495

In order to study the robustness of the features extraction technique the Kth nearest classification technique is also applied on the same data set. And the results for the two different set of different K values are represented in two parts i.e. results with K = 1 and the results with K = 3. The confusion tables for the K=1 for steerable filter technique are represented in Tables 6-to 9.

TABLE 6: CONFUSION TABLE OF ENGLISH AND TAMIL SCRIPT CLASSIFICATION WITH KNN WHEN K=1

Scripts	English	Tamil
English	486	14
Tamil	10	490

TABLE 7: CONFUSION TABLE OF ENGLISH AND URDU SCRIPT CLASSIFICATION WITH KNN WHEN K=1

Scripts	English	Urdu
English	500	0
Urdu	0	500

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

TABLE 8: CONFUSION TABLE OF ENGLISH AND HINDI SCRIPT CLASSIFICATION WITH KNN WHEN K=1

Scripts	English	Hindi
English	484	16
Hindi	16	484

TABLE 9: CONFUSION TABLE OF ENGLISH AND KANNADA SCRIPT CLASSIFICATION WITH KNN WHEN K=1

Scripts	English	Kannada
English	478	12
Kannada	16	484

The results for steerable filter method with Kth Nearest neighbor tech with K value set as 3 are represented in the Table 10. Also the confusion tables for the K=3 values are represented in Tables 10 to 14.

TABLE 10: BI-SCRIPT CLASSIFICATION RESULTS WITH KNN (K=3) AND 10 FOLD CROSS VALIDATION

Script	Kannada	Hindi	Urdu	Tamil	Average
English	95.0	92.0	100.0	96.0	95.75

TABLE 11: CONFUSION TABLE OF ENGLISH AND KANNADA SCRIPT CLASSIFICATION WITH KNN WHEN K=3

Scripts	English	Kannada
English	484	16
Kannada	20	480

TABLE 12: CONFUSION TABLE OF ENGLISH AND HINDI SCRIPT CLASSIFICATION WITH KNN WHEN K=3

Scripts	English	Hindi
English	482	18
Hindi	24	476

TABLE 13: CONFUSION TABLE OF ENGLISH AND URDU SCRIPT CLASSIFICATION WITH KNN WHEN K=3

Scripts	English	Urdu
English	500	0
Urdu	0	500

TABLE 14: CONFUSION TABLE OF ENGLISH AND TAMIL SCRIPT CLASSIFICATION WITH KNN WHEN K=3

Scripts	English	Tamil
English	482	18
Tamil	20	480

From the above results presented it is evident that the Steerable Gaussian filtering techniques based features are robust in terms of accuracy as the accuracy of the same is found to be 99.125% with Linear Discriminate classification technique.

VI. CONCLUSION

In this work, a method for printed word-wise script identification based on steerable Gaussian filtering technique is proposed. The results are analysed using two different classification techniques to study the robustness of the features extracted as well as the properties and structural shape differences of 5 Indian scripts. The results are found to be encouraging with both classification methods. The Linear discriminate classifier gives an accuracy of 99.125%. The KNN classifier provided the accuracy of about 96% and the optimal value of K was found to be K = 1. It is observed that every script has a distinct visual and textural appearance. The structural shape properties of Indian scripts are direction sensitive. Hence, to model these direction sensitive properties, Steerable filters are well suited which is revealed from the efficacy of the method in modelling the properties of Human Visual System. Exhaustive experimentation was carried out at all levels and it is observed that the results obtained are better in terms of accuracy and are comparable with the existing methods.

REFERENCES

[1] D.P.Pattanayak, "Literacy Education," K. Rajyashree, et al. Eds. Mysore: Central Institute of Indian Languages, 1980.
[2] J.C.Shrama, "Language and script in India: Some Challenges," http://www.languageinindia.com/sep2001/jcscript.html/2001
[3] U. Pal and B. B. Chaudhuri, "Script line separation from Indian multi-script documents," IETE Journal of Research, vol. 49, pp. 3-11, 2003.
[4] U. Pal, S. Sinha, and B. B. Chaudhuri, "Multi-Script Line identification from Indian Documents," In Proceedings of 7th International Conference on Document Analysis and Recognition, pp. 880-884, 2003.
[5] U. Pal and B. B. Chaudhuri, "Automatic Identification Of English, Chinese, Arabic, Devnagari And Bangla Script Line," In Proc of Sixth International Conference on Document Analysis and Recognition, pp. 790-794, 2001.
[6] U. Pal and B. B. Chaudhuri, "Identification of different script lines from multi-script documents," Image and Vision computing, vol. 20, pp. 945-954, 2002.
[7] R. Manthalkar and P. K. Biswas, "An Automatic script identification scheme for Indian Languages," www.ee.iitb.ac.in/uma/~ncc2002/proc/NCC-02/pdf/n028 .ps.
[8] G. D. Joshi, S. Garg, and J. Sivaswamy, "Script Identification for Indian Documents," In Proceedings of 7th IAPR workshop on Document Image Systems (DIS), New Zealand, pp. 255-267, 2006.
[9] M. C. Padma and P. Nagabhushan, "Horizontal and vertical linear edge features as clues in the discrimination of multilingual (Kannada, Hindi and English) Machine printed documents," In Proceedings of National Workshop on Computer Vision, Graphics and Image Processing- WVGIP, pp. 204-209, 2002.
[10] M. C. Padma and P. Nagabhushan, "Identification and separation of text words of Kannada, Hindi and English languages through Discriminating features," In Proceedings of 2nd National Conference on Document Analysis and Recognition, pp. 252-260, 2003.
[11] M.C.Padma and P.A.Vijaya, "Identification and Separation of Text Words of Kannada, Telugu, Tamil, Hindi, English Languages through Visiual Discriminating Features," In Procs of International Conference on Advances in Computer Vision and Information Technology, pp. 1283-91, 2007.
[12] S. B. Patil and N. V. Subbareddy, "Neural network based system for

script identification in Indian documents," Sadhana, vol. 27, pp. 83-97, 2002.
[13] D. Dhanya, A. G. Ramakrishna, and P. B.Pati, "Script identification in printed bilingual documents," Sadhana, vol. 27, pp. 73-82, 2002.
[14] D. Dhanya and A. G. Ramakrishna, "Script identification in printed bilingual documents," In Proceedings of Document Analysis and Systems, pp. 13-24, 2002.
[15] 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.
[16] P. B. Pati and A.G.Ramakrishnan, "HVS inspired system for Script Identification in Indian Multi-Script Documents," In Proceedings of 7th International Workshop on Document Analysis System, Nelson Newland, pp. 380-389, 2006.
[17] P. Pati and A. Ramakrishna, "A Blind Indic Script Recognizer for Multi-script Documents," In Proceedings of Ninth International Conference on Document Analysis and Recognition, vol. 2, pp. 1248-1252, 2007
[18] B.B.Chaudhuri and U.Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi)," In Proceedings of Fourth International Conference on Document Analysis and System, pp. 1011-1016, 1997.
[19] T. N. Vikram and D. S. Guru, "Appearance based models in document script identification," In Proceedings of ICDAR 2007, vol. 2, pp. 709-713, 2007.
[20] P. B. Pati and A. G. Ramakrishna, "Word level multi-script identification," Pattern Recognition Letters, vol. 29, pp. 1218–1229, 2008.
[21] U. Pal and B. B. Chaudhuri, "Script line separation from Indian Multi-script documents," In Proceeding 5th International Conference on Document Analysis and Recognition, pp. 406-409, 1999.
[22] B V Dhandra et. al, Script Identification based on Morphological Reconstruction in Document Images", In Proc. of 18th International Conference on Pattern Recognition (ICPR2006) Hong Kong, Aug. 2006, Vol. II-3, pp 950-53.
[23] B V Dhandra et. al, Word wise script identification in bilingual documents based on Morphological reconstruction", In Proc. of 1st IEEE International Conference on Digital Image Management (ICDIM2006), Bangalore India, held during 6-8 Dec. 2006, pp 389-94.
[24] B V Dhandra et. al, Word- wise Script Identification based on Morphological Reconstruction in printed Bilingual Documents", In Proc. of IET International Conference on Visual Information Engineering (VIE2006) Bangalore, 28-29 Sept. 2006, pp 389-393. .
[25] 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.
[26] W. Freeman and E. Adelson. "The design and use of steerable filters," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, no. 9 pp. 891-906, 1991.
[27] Ghosh D, T. Dube, and A.P. Shivaprasad Script Recognition – A Review IEEE PAMI, Vol. XX, no. YY, 2010 pp 1-21.
[28] Pan W. M., C. Y. Suen, T. D. Bui, Script Identification Using Steerable Gabor Filters, Proceedings of the Eight International Conference on Document Analysis and Recognition (ICDAR'05) 2005.
[29] Rafael Gonzalez and R Woods, IInd Ed., Digital Image Processing, Pearson Education, 2004.
[30] Anil K.Jain, "Fundamentals of Digital Image Processing", PHI, 2006.

BIOGRAPHIES

Dr. Virendra. S. Malemath is a faculty member in Department of Computer Science & Engg, KLEDR M S Sheshgiri College of Engg. & Tech., Belgaum. He did his Bachelors in Engg. in Electronics & Communication Engg. from Karnataka University, Dharwad in the year 1993, did his MS in Software Systems from BITS Pilani Rajasthan in 1998 and received his PhD in Computer Science from Gulbarga University, Gulbarga, India in 2009. His research interests are document image processing and pattern recognition. He has published more than 40 articles in peer reviewed international journals and conferences.

Prof. A H Kulkarni is a senior Faculty in the Department of Computer Science & Engg KLS Gogte Institute of Technology, Belgaum. He did his Bachelors in Engineering in Comp. Sc. & Engg. from Basaveshwar Engineering. College, Bagalkot in the year 1996, did his Masters of Technology in Computer Science & Engg. from Visvesvaraya Technological University, Belgaum in 2001. His research interests include Document image processing, Medical Image Processing and pattern recognition. He has published more than 25 articles in peer reviewed international journals and conferences.

Dr. H. Mallikarjun is the Head of Department of PG Studies and Research in Computer Science and coordinator of IQAC at KASCC, Bidar, India. He has received his master degree in Statistics from Gulbarga University, in 1989. He has completed his M. Phil in Computer Science from M.S University, Tamil Nadu, India in 2003. He has received his PhD in Computer Science from Gulbarga University, Gulbarga, India, in 2009. His research interests are image processing and pattern recognition. He has published 40 articles in peer reviewed international journals and conferences.