Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

K-means Clustering for Document Analysis in Indian Bilingual Documents

[1]Mahesha D M, [2]Chethan H K, [3]Gopalan N P
[1] Department of Studies in Computer Science, Karnataka State Open University, Mysore, India
[2] Department of Computer Science and Engineering, Maharaja Institute of Technology, Mysore, India
[3]Department of Computer Applications NIT, Trichirappalli, India

Abstract*— Script identification is one of the preprocessing steps in any document image processing task. Script identification in printed documents has achieved a greater attention. In this paper, we present a scheme to identify different Indian scripts from a document image. For given script we extracted different features like Gray Level Co-occurrence Method (GLCM) and Local Binary Pattern (LBP) features. The features are extracted globally from a given text block which does not require any complex and reliable segmentation of the document image into lines and characters. The features are unsupervised classified using K-means clustering. The scheme has been tested on 200 Indian scripts and found to be robust in the process of scanning and relatively insensitive to change in font size. This proposed system achieves good classification accuracy on a large testing data set.

***Keywords*— GLCM, LBP, K-means**

I. INTRODUCTION

Document image analysis has been an active research area from a few decades, and that facilitates the establishment of paperless offices across the world. The process of converting textual symbols present on printed and/ or handwritten paper to a machine understandable format is known as optical character recognition (OCR) which is the core of the field of document image analysis. The OCR technology for Indian documents is in emerging stage and most of these Indian OCR systems can read the documents written in only a single script. As per the trilingual formula of Indian constitution [1], every state Government has to produce an official document containing a national language (Hindi), official language (English) and state language (or regional language).

According to the three-language policy adopted by most of the Indian states, the documents produced in an Indian state Karnataka, are composed of texts in the regional language-Kannada, the National language-Hindi and the world wide commonly used language-English. In addition, majority of the documents found in most of the private and Government sectors of Indian states, are tri-lingual type (a document having text in three languages). So, there is a growing demand to automatically process these tri-lingual documents in every state in India, including Karnataka.

The monolingual OCR systems will not process such multi-script documents without human involvement for delineating different script zones of multi-lingual pages before activating the script specific OCR engine. The need for such manual involvement can result in greater expense and crucially delays the overall image-to-text conversion. Thus, an automatic forwarding is required for the incoming document images to handover this to the particular OCR engine depending on the knowledge of the intrinsic scripts. In view of this, identification of script and/ or language is one of the elementary tasks for multi-script document processing. A script recognizer, therefore, simplifies the task of OCR by enhancing the accuracy of recognition and reducing the computational complexity.

II. PREVIOUS WORK

Existing works on automatic script identification are classified into either local approach or global approach. Local approaches extract the features from a list of connected components like line, word and character in the document images and hence they are well suited to the documents where the script type differs at line or word level. In contrast, global approaches employ analysis of regions comprising of at least two lines and hence do not require fine segmentation. Global approaches are applicable to those documents where the whole document or paragraph or a set of text lines is in one script only. The script identification task is simplified and performed faster with the global rather than the local approach. A sample work has been reported in literature on both Indian and non-Indian scripts using local and global approaches.

A. Local approaches on Indian scripts

Pal and Choudhuri [2] have proposed an automatic technique of separating the text lines from 12 Indian scripts (English, Hindi, Bangla, Gujarati, Tamil, Kashmiri, Malayalam, Oriya, Punjabi, Telugu and Urdu) using ten triplets formed by grouping English and Devanagari with any one of the other scripts. This method works only when the triplet type of the document is known. Script identification technique explored by Pal [3] uses a binary tree classifier for 12 Indian scripts

using a large set of features. B Patil and Subbareddy [4] have proposed a neural network based system for script identification of Kannada, Hindi and English languages. Dhandra et al., [5] have exploited the use of discriminating features (aspect ratio, strokes, eccentricity, etc,) as a tool for determining the script at word level in a bi-lingual document containing Kannada, Tamil and Devnagari containing English numerals. A method to automatically separate text lines of Roman, Devanagari and Telugu scripts has been proposed by Pal et al., [6]. In Lijun et al, [7] have developed a method for Bangla and English script identification based on the analysis of connected component profiles. Vipin [8] have presented an approach to automatically identify Kannada, Hindi and English languages using a set of features viz., cavity analysis, end point analysis, corner point analysis, line based analysis and Kannada base character analysis. Word-wise script identification systems for Indian scripts has been discussed in [22].

B. Global approaches on Indian scripts

Adequate amount of work has been reported in literature using global approaches. S Chaudhury et al., [9] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. G D Joshi et al., [10] have presented a script identification technique for 10 Indian scripts using a set of features extracted from logGabor filters. Dhanya et al., [11] have used Linear Support Vector Machine (LSVM), K-Nearest Neighbour (K-NN) and Neural Network (NN) classifiers on Gabor-based and zoning features to classify Tamil and English scripts. Hiremath [12] have proposed a novel approach for script identification of South Indian scripts using wavelet based co-occurrence histogram features. S R Kunte and S Samuel [13] have suggested a neural approach in on-line script recognition for Telugu language employing wavelet features. Peeta et al., [14] have presented a technique using Gabor filters for script identification of Indian bilingual documents.

C. Local and global approaches on non-Indian scripts

Sufficient amount of work has also been carried out on non-Indian languages. Spitz [15] has proposed a system, which relies on specific, well defined pixel structures for script identification. Such features include locations and numbers of upward concavities in the script image, optical density of connected components, the frequency and combination of relative character heights. This approach has been shown to be successful in distinguishing between Asian languages (Japanese, Chinese, and Korean) against European languages (English, French, German, and Russian). Wood et al., [16] have proposed projection profile method to determine Roman, Russian, Arabic, Korean and Chinese characters. Hochberg et al., [17] have presented a method for automatically identifying script from a binary document image using cluster-based text symbol templates. In Ding et al., [18], a method that uses a combined analysis of several discriminating statistical features to classify Oriental and European scripts is presented. Tan et al., [19] has proposed a rotation invariant texture feature extraction method for automatic script and language identification from document images using multiple channel (Gabor) filters and Gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Koreans, Malayalam, Persian and Russian. A Busch et al., [20] has presented the use of texture features (gray level co-occurrence matrix and Gabor energy features) for determining the script of a document image. B.Kumar et al. [21] have used topological, structural features with rule based classifier for line based multi-script identification.

It can be seen from the references cited above that sample amount of work has been done in the area of document script/language identification. Even though some considerable amount of work has been carried out on Indian script identification, hardly few attempts focus on the all the languages. So, an intensive work needs to be done in this field as the demand is increasing. Also the existing methods have to be improved to reach a stage of satisfactory practical application. It is in this direction the research work proposes a model that automatically identifies the all the languages in given document. We propose a based classification scheme which uses a global approach. In section (2), we describe the proposed scheme in detail. Results of the scheme tested over a large data set are presented in section (3).

III. PROPOSED METHOD

The steps of proposed method are segmentation, Features extraction, Classification.

A. Segmentation

Our scheme first segments the text area from the document image by removing the upper, lower, left and right blank regions. After this stage, we have an image which has textual and non-textual regions. This is then binarised after removing the graphics and pictures. It should be noted that the text block may contain lines with different font sizes and variable spaces between lines words and characters.

B. Feature Extraction

Feature extraction is a necessary step for any classification task. For image object classification purpose, the use of texture features has proved to be quite effective for many applications. There are many ways for calculating texture feature descriptors. We choose GLCM and LBP as feature descriptors, since both have been used successfully for object recognition, robotic mapping etc.

1) GLCM Based Texture Feature Descriptors

Texture features based on spatial co-occurrence of pixel values are probably the most widely used texture feature descriptors having been used in several application domains like analysis of remotely sensed images, image segmentation etc. Cooccurrence texture features are extracted from an image into two steps. First, pair wise spatial co-occurrence of

pixels separated by a given angular value are computed and stored in a grey level co-occurrence matrix. Second, the GLCM is used to compute a set of scalar quantities that characterizes the different aspects of the underlying texture. We have worked with four GLCM based descriptors, namely, Contrast, Correlation, Homogeneity and Energy [26].

2) Local Binary Pattern

Ojala et al. [27] proposed to use the Local Binary Pattern (LBP) histogram for rotation invariant texture classification. LBP is a simple but efficient operator to describe local image patterns. It is combined statistical and structured method.

LBP is a gray-scale texture operator that characterizes the local spatial structure of the image texture. The basic LBP operator considers a 3x3 neighborhood of a pixel, then these 8 border pixels will be replaced either by 1, if they are larger than or equal to the central pixel or by 0 otherwise. Finally, the central pixel will be replaced with a summation of the binary weights of border pixels in the LBP image and the 3x3 window slides to the next pixel.

It is possible to develop the basic LBP into various neighborhood sizes and distances.

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(g_P - g_c)2^P \qquad (1)$$

Where *s* (.) is the sign function:

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (2)$$

g_p and g_c are grey levels of border pixels and central pixel respectively, *P* is the number of pixels in the neighborhood and R is the radius of the neighborhood. Suppose the coordinates of g_c are (0, 0), then the coordinates of g_p are given by (-R sin (2πp/P), R cos (2πp/P)).

In this case, if we set (*P* = 8; *R* = 1), we obtain the basic LBP (1) Luminance changing cannot affect signed differences g_p-g_c, hence LBP is grey level shift invariant. Suppose the texture image is NxM. After identifying the LBP pattern of each pixel (i, j), the whole texture image is represented by building a histogram:

$$H(k) = \sum_{i=1}^{N} \sum_{j=1}^{M} f(LBP_{P,R}(i,j), k), k \in [0, K] \qquad (3)$$

$$f(x,y) = \begin{cases} 1, & x = y \\ 0, & otherwise \end{cases} \qquad (4)$$

where K is the maximal LBP pattern value. The U value of an LBP pattern is defined as the number of spatial transitions (bitwise0/1 changes) in that pattern

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{P=1}^{P-1} |s(g_P - g_c) - s(g_{P-1} - g_c)| \qquad (5)$$

For example LBP pattern 00000000 has a U value of 0 and 01000000 of 2.The uniform LBP pattern refers to the uniform appearance pattern which has limited transition or discontinuities (U≤2) in the circular binary presentation. It was verified hat only ‘‘uniform’’ patterns are fundamental patterns of local image texture. In practice, the mapping from $LBP_{P,R}$ to $LBP^{u2}_{P,R}$ (superscript “*u2*” means uniform patterns with *U*≤2),which has *P**(*P*-1)+3 distinct output values, is implemented with a lookup table of 2^P elements.

To achieve rotation invariance, a locally rotation invariant pattern could be defined as:

$$LBP^{riu^2}_{P,R} = \begin{cases} \sum_{P=0}^{P-1} s(g_P - g_c) & if\ U(LBP_{P,R}) \leq 2 \\ P+1 & otherwise \end{cases} \qquad (6)$$

The mapping from $LBP_{P,R}$ to $LBP^{riu^2}_{P,R}$ (superscript “*riu2*”means rotation invariant “uniform” patterns with *U*≤2), which has *P*+2 distinct output values, can be implemented with a lookup table.

IV. CLASSIFICATION

K-Means algorithm is an unsupervised classification technique, where the user initiates the algorithm by specifying the number of clusters to be created from feature sets of an image [28]. This algorithm splits the given image into different clusters of features in the feature space, each of them defined by its center. Initially each feature in the image is allocated to the nearest cluster. Then the new centers are computed with the new clusters. These steps are repeated until convergence. Basically we need to determine the number of clusters K first. Then the centroid will be assumed for these clusters. We could assume random objects as the initial centroids or the first K objects in sequence could also serve as the initial centroids. In the proposed method we consider two clusters because only two classes are required. One is kannada and another one is English.

V. EXPERIMENTATION

In order to conduct experimentation, we have considered three classes i.e., bi lingual scripts Kannada and English. For each class 100 documents were created.

In this work, we intend to study the classification accuracy under varying features of PCA. We pick samples randomly from the database and experimentation is conducted on database of more than 100 samples. The Figure 1 shows accuracy using individual features like LBP, GLCM under varying reduction PCA features from 10% to till results is saturated. The experimentation is conducted more than five times and best one is picked from two classes. From Fig3 we can understand that the GLCM features are normalized at 83.55 by 60% reductions of features, LBP features are normalized at 87 by 60% reductions of features.

VI. CONCLUSION

In this work, we present a method to identify different Indian scripts from a document image. For given script we extracted different features like Gray Level Co-occurrence Method (GLCM) and Local Binary Pattern (LBP) features.

The features are extracted globally from a given text block which does not require any complex and reliable segmentation of the document image into lines and characters. The features are unsupervised classified using K-means clustering. The scheme has been tested on 200 Indian scripts and found to be robust in the process of scanning and relatively insensitive to change in font size.

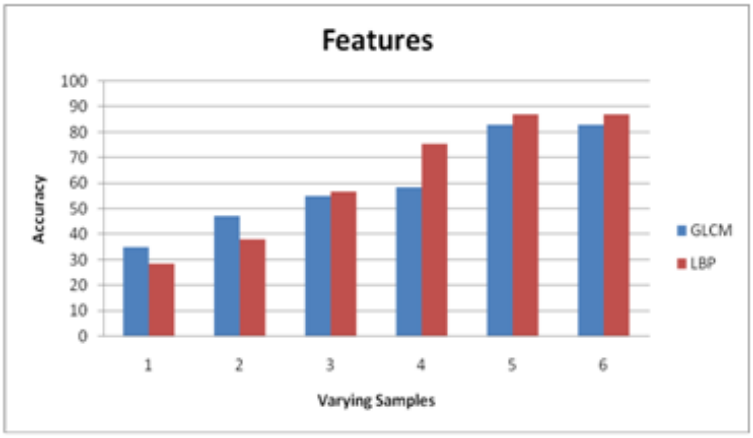

Fig 1: Graphical Representation of Individual Feature (LBP, GLCM)

REFERENCES

[1] Pal U, S. Sinha and B. B.Chaudhri, "Multi-Script Line Identification from Indian Documents", Proceedings of International Conference on Document Analysis and Recognition, (2003) August, pp. 880-884.

[2] Pal U., Chaudhuri B.B., (1999), Script line separation from Indian multi-script document, Proc. 5th Int. Conf. on Document Analysis and Recognition (IEEE Comput. Soc. Press), 406–409.

[3] Pal U. and Chaudhuri B.B., (2003), Script line identification from Multi script documents, IETE journal Vol. 49, No 1, 3-11.

[4] Basavaraj Patil S. and Subbareddy N.V., (2002), Neural network based system for script identification in Indian documents, Sadhana Vol. 27, Part 1, 83–97.

[5] Dhandra B.V., Nagabhushan P., Mallikarjun Hangarge, Ravindra Hegadi, Malemath V.S., (2006), Script Identification Based on Morphological Reconstruction in Document Images, The 18th International Conference on Pattern Recognition (ICPR'06), Vol.No. 11-3, 950-953.

[6] Pal U., Chaudhuri B.B., (1999), Script line separation from Indian multi-script document, Proc. 5th Int. Conf. on Document Analysis and Recognition (IEEE Comput. Soc. Press), 406–409.

[7] Lijun Zhou, Yue Lu and Chew Lim Tan, (2006), Bangla/English Script Identification based on Analysis of Connected component Profiles, Proc. 7th IAPR workshop on Document Analysis System, New land, 234-254.

[8] Vipin Gupta, G.N. Rathna, K.R. Ramakrishnan,: A Novel Approach to Automatic Identification of Kannada, English and Hindi Words from a Trilingual Document, Int. conf. on Signal and Image Processing, Hubli, pp. 561-566, (2006).

[9] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, Shet R.B., (2000), Identification of scripts of Indian languages by Combining trainable classifiers", Proc. of ICVGIP, India.

[10] Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy, (2006), Script Identification from Indian Documents, H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, 255–267.

[11] Dhanya D., Ramakrishnan A.G. and Pati P.B., (2002), Script identification in printed bilingual documents, Sadhana, vol. 27, 73-82.

[12] Hiremath P S and S Shivashankar, "Wavelet Based Co-occurrence Histogram Features for Texture Classification with an Application to Script Identification in a Document Image", Pattern Recognition Letters 29, 2008, pp 1182-1189.

[13] Srinivas Rao Kunte R. and Sudhakar Samuel R.D., (2002), A Neural Approach in On-line Script Recognition for Telugu Language Employing Wavelet Features, National Workshop on Computer Vision, Graphics and Image Processing (WVGIP), 188-191.

[14] Peeta Basa Pati, S. Sabari Raju, Nishikanta Pati and A. G. Ramakrishnan, "Gabor filters for Document analysis in Indian Bilingual Documents", 0-7803-8243-9/04/ IEEE, IClSlP, pp. 123- 126, 2004.

[15] Spitz A. L., (1994), Script and language determination from document images, Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 229-235.

[16] Wood S. L.; Yao X.; Krishnamurthy K. and Dang L., (1995): Language identification for printed text independent of segmentation, Proc. Int. Conf. on Image Processing, 428–431, IEEE 0- 8186-7310-9/95.

[17] Hochberg J., Kerns L., Kelly P. and Thomas T., (1997), Automatic script identification from images using cluster based templates, IEEE Trans. Pattern Anal. Machine Intell. Vol. 19, No. 2, 176–181.

[18] Ding J., Lam L. and Suen C. Y., (1997), Classification of oriental and European Scripts by using Characteristic features, Proc. 4th ICDAR , 1023-1027.

[19] Tan T. N., (1998): Rotation invariant texture features and their use in automatic script identification, IEEE Trans. Pattern Anal. Machine Intell. PAMI, Vol.20, No. 7, 751–756.

[20] Andrew Busch; Wageeh W. Boles and Sridha Sridharan, (2005), Texture for Script Identification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 11, pp. 1720-1732.

[21] B. Kumar, A. Bera and T. Patnaik, "Line Based Robust Script Identification for Indian Languages", International Journal of Information and Electronics Engineering, vol. 2, no. 2 (2012), pp. 189-192.

[22] R. Rani, R. Dhir and G. S. Lehal, "Modified Gabor Feature Extraction Method for Word Level Script Identification- Experimentation with Gurumukhi and English Scripts", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6, no. 5, (2013), pp. 25-38.

[23] A. K. Jain and Y. Zhong., Page segmentation using texture analysis. Pattern Recognition 29 (1996) 743–770.

[24] http://www.samachar.com/.

[25] Digital Library of India. http://dli.iiit.ac.in/

[26] R. M. Haralick, K. Shanmugam, and I. Dinstein, Textural Features of % Image Classification, IEEE Transactions on Systems, Man and Cybernetics, % vol. SMC-3, no. 6, Nov. 1973.

[27] T. Ojala, M. Pietikäinen, and T. T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with Local Binary Pattern, IEEE Trans. on PAMI 24(7), pp. 971-987, 2002.

[28] K. Somasundaram1 and N. Kalaichelvi, Feature Identification in Satellite Images using K-Means Segmentation, NCSIP-2012, pp.264-269 .

