Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 3, No. 2, March : 2023

SCRIPT IDENTIFICATION USING HYBRID FEATURES AND SVM CLASSIFIER

Dr. Shailesh A. Chaudhari, Assistant Professor, Department of ICT, Veer Narmad South Gujarat university, Surat (Gujarat), India, sachaudhari@vnsgu.ac.in

Abstract

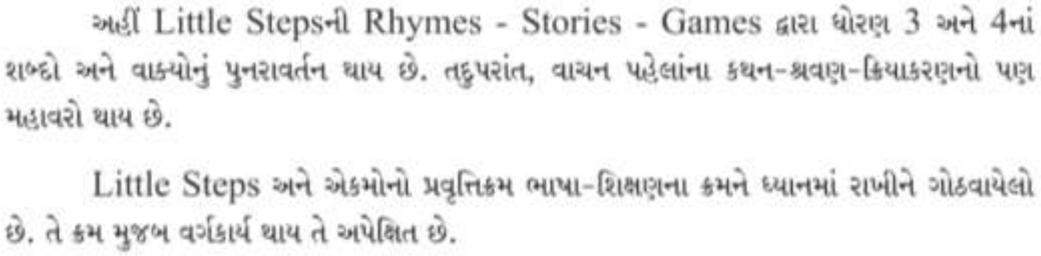
Script identification is a difficult issue in a multilingual optical character recognition. In either an Indian or non-Indian environment, a notable research work on script identification has been highlighted. There are a lot of bilingual commercial and official regional documents from the many Indian states, with English serving as the international intersperse language. Thus, one of the main tasks in the recognition of multi-script documents is script identification. Regional documents from various Indian states frequently include English terminology. This study presents Gujarati and English word-level script identification. Word images are used to obtain statistical features for feature extraction. Zone-based pixel density is computed here. The suggested method classifies the extracted features in one of the scripts using an SVM classifier with various kernel functions. According to the experiment, SVM polynomial function performs better compared to other SVM kernel functions.

Keywords:
Script Identification, Classification, Support Vector Machine, Feature Extraction.

1. INTRODUCTION

Researchers have been working on optical character recognition (OCR) for a few decades, and a lot of their work can be found in the literature. When the document to be identified is monolingual, OCR is easy; however, when the material is bilingual or multilingual, OCR is challenging. Several regional publications, periodicals, reports, and documents are bilingual in India and contain both English and the native language of the region. For OCR researchers, processing this kind of material presents a difficult task. By increasing recognition accuracy, a script identifier makes OCR easier.

There are 12 different scripts used to write the 18 official Indian languages, although English has traditionally been used as the nation's legal language. Recognition of Indian script is still in its infancy. In India, almost all official documents are issued by the government and frequently include both English and other official dialects. Fig. 1 illustrates how English terms are incorporated into a Gujarati text from the state of Gujarat. Word level script identification is therefore necessary for an OCR system to function on these types of texts.



અહીં Little Stepsની Rhymes - Stories - Games દ્વારા ધોરણ 3 અને 4નાં શબ્દો અને વાક્યોનું પુનરાવર્તન થાય છે. તદુપરાંત, વાચન પહેલાંના કથન-શ્રવણ-ક્રિયાકરણનો પણ મહાવરો થાય છે.

Little Steps અને એકમોનો પ્રવૃત્તિક્રમ ભાષા-શિક્ષણના ક્રમને ધ્યાનમાં રાખીને ગોઠવાયેલો છે. તે ક્રમ મુજબ વર્ગકાર્ય થાય તે અપેક્ષિત છે.

Fig. 1. bi-script document image of Gujarati and English words

An approach to word-level script identification for printed bilingual (Gujarat-English) documents is suggested in this research. Statistical and structural aspects are both utilised in this context. It is explored how SVM with various kernel functions compare in terms of classifier accuracy. Experiments are conducted in the training and testing datasets to investigate the relationship between font type and font size.

This is how the paper is structured: The Section 2 describes related work. The feature extraction algorithms are described in Section 3. Section 4 goes into detail on the SVM Classifier. Sections 5 and 6 present the conclusion and potential future improvements after discussing the experimental analysis and outcomes.

2. RELATED WORK

The blending of various scripts at the paragraph, line, word, or character levels inside a single document is described in [1,2]. The smallest size of the text that may be consistently used to extract features determines the employment of a script identification system. To distinguish a script from a multilingual document, there are two basic methods: local approach and global approach. The global approach analyses areas with at least two lines and does not call for any kind of precise segmentation. While the local technique analyses related components seen in document images, and such components necessitate image segmentation as a separate processing step.

Major contributions to the field of Indian language document analysis are made by Chaudhari and Pal [3, 4, 5, 12]. They used the headline as a characteristic to set English lines apart from the Bangla and Devanagari lines. In order to distinguish between other Indian languages, they also used a variety of structural characteristics.

Padma and Vijaya [6] presented a texture-based methodology. They use these features to identify the script in a document that was mechanically printed by utilising Wavelet Packet Decomposition to derive the Haralick texture characteristics from the co-occurrence matrix. Wavelet packets decompose each detail coefficients vector in the same way as approximation vectors do.

Block level script identification is done in two steps, according to Patil and Subbareddy [7]. Using 3X3 matrix masks, the document image is first dilated in the horizontal, vertical, right diagonal, and left diagonal directions. The four modified versions of the original image are used to construct a feature vector in the second stage. An appearance-based paradigm for script identification at the paragraph and word levels was proposed by Vikram and Guru [9]. They employed appearance-based models based on FLD (Fisher's Linear Discriminator) and PCA (Principal Component Analysis) for feature extraction.

To differentiate between the Tamil and English scripts, Dhanya et al.'s[10] initial word-level script identification effort. Character density, word-level Gabor filters with suitable frequencies and orientations, and the horizontal spatial dispersion projection profiles of a word in the upper and lower zones were employed for feature extraction. An SVM (Support Vector Machine) based technique was proposed by Chanda et al. [11] for word level script detection from Indian papers containing Bengali, Devanagari, and English. The 64-dimensional chain-code characteristics they developed for categorization. The samples that were rejected at the initial classification level are further processed using the 400-dimensional gradient feature before being classified at the subsequent level.

For script recognition at the word level, Dhandra et al. [8,15,16] developed a method based on morphological reconstruction. They employed morphological erosion and opening to reassemble words in the four directions of the horizontal, vertical, right, and left diagonal. They also filled in gaps in loop-containing characters. Kunte et al. [13] presented a Gabor feature extraction technique that may be used with a pre-trained neural classifier to recognize the script at the word level.

An OCR system that can recognize Tamil and English scripts at the character level and is font and size independent was presented by Aparna et al[14].They used elements like discrete cosine transform coefficients and geometric moments. First effort for bilingual printed Gujarati-English documents was reported by Chaudhari and Gulati. In printed bilingual Gujarati-English papers, statistical features were used for script recognition at the line level [18].

For bilingual scripts, the unified technique built on the SFTA feature set created by Dhandra B. V. et al. performs considerably better [19]. They claimed that, with recognition accuracy ranging from 62% to 82.50%, geometric characteristics (10 features) only exceed SFTA features (nt=8) in differentiating Gujarati with English and Punjabi with English scripts. Mahajan S. and Rani R. presented a method for distinguishing a script from a genuine scene[20]. They experimented on four datasets and introduced CNN and the Inception network to identify the script. Comparatively, the competitive outcomes have been attained. Their approach can address the important problems brought on by the camera, the image, and the background.

3. FEATURE EXTRACTION

Any character recognition system must consider feature extraction. The goal of feature extraction is to characterize a pattern using the fewest information possible while still being able to distinguish between distinct classes of patterns. For feature extraction in this study, statistical and structural features are used.

A collection of features are taken from the image before moving on to the recognition stage. These elements provide a compact representation of the image's contents that are crucial to the process of recognition. In order to create a vector of values that will be provided to the classifier, the retrieved features are further processed.

Statistical feature and structural feature are two of the sorts of methodologies utilized in feature extraction.

3.1 STATISTICAL FEATURE EXTRACTION

Style variations in document images are somewhat handled by this type of features. They primarily serve to shrink the size of the feature set, resulting in fast speed and minimal complexity. There are three distinct zones in which the words of both the Gujarati and English scripts can be divided. Gujarati and English script's upper, middle, and lower zones are depicted in Fig. 2.

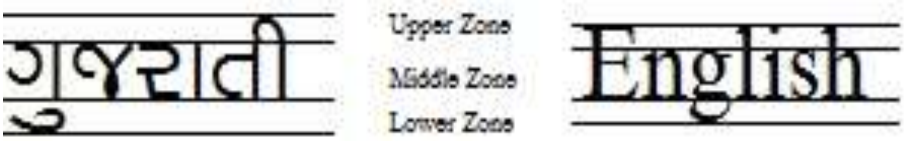


Fig. 2. Zones of Gujarati and English Script

From each word image, two structural features and nine different statistical features are extracted.

3.1.1 Upper Zone Pixel Density:

This is derived by dividing the upper zone's total number of pixels by its area.

3.1.2 Lower Zone Pixel Density:

It is derived by dividing the size of the lower zone by the sum of the pixels in the lower zone.

3.1.3 Middle Zone Pixel Density:

This metric measures how many pixels are present in the middle zone relative to its size.

3.1.4 Character density:

IT is computed by dividing the word's height by the total number of characters.

3.1.5 Vertical Line:

Vertical line present in the many characters. It shows two groups that are separated into independent and connected vertical lines.

3.1.6 Horizontal Line:

Characters are made up of horizontal lines. It stands for a particular set of these characters.

3.1.7 A positive diagonal line :

It denotes a character with a positive slope.

3.1.8 A negative diagonal line:

It denotes a character with a negative slope.

3.1.9 Close Region (Loop):

Loop denotes any near region in a character, and based on how many close regions a character has, it is further divided into two groups.

3.2 STRUCTURAL FEATURE EXTRACTION

These features are based on the character's topological and geometrical characteristics. Geometrical and topological features with a variety of global and local qualities can represent the characters.

3.2.1 *End Point:*

The end point defines the character's beginning and ending marks. For classification, the end points of each sub component of a character with disconnected components are taken into consideration.

3.2.2 *Cross Point:*

A cross point is an intersection point or a midpoint where two structures cross.

4. CLASSIFICATION

The main objective of classification is to categorize an object or pattern using the feature vectors that the feature extraction technique provides. A thorough investigation has been completed through experimental tests utilizing SVM classifier on bi-script databases to investigate the behaviour of the suggested technique.

4. 1. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM), one of the most popular techniques for supervised learning, is used to address Classification and Regression problems. Machine learning typically uses it to address classification problems. It was made by Cortes and Vapnik [17] in the early months of 1946.

The strength of SVM lies in its ability to convert data into a high-dimensional space that can be partitioned using a hyperplane. SVM is a well-established method for producing an ideal hyperplane that distinguishes between two classes by increasing the distance or margin between the two classes. Consequently, it is important to comprehend the optimization method for SVM learning with various hyper plane parameters.

SVM categorizes data points even when they are not linearly separable by mapping the data to a high-dimensional feature space. Once a separator between the categories is identified, the data are transformed to enable the hyperplane representation of the separator. The group to which a new data point should belong can therefore be predicted using the features of new data point.

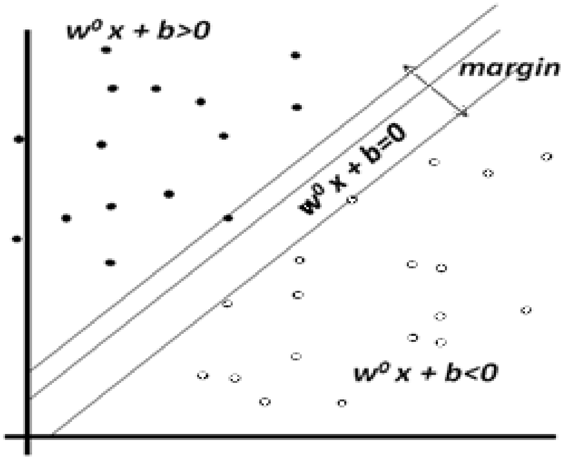


Fig. 3.Two class support vector

The hyperplane can be represented as in Eq. 1 according to Figure 3. (1)

$$w^0 . x + b = 0 \tag{1}$$

In SVM, a kernel function is utilized to aid in problem solving. We can carry out calculations smoothly on high dimensional data with the kernel's assistance. Kernels allow us to expand to an endless number of dimensions. The classification and analysis of specific data set patterns rely heavily on the kernel. They are extremely valuable when employing a linear classifier to solve a non-linear issue.

Some commonly used kernel types are:

1.) Linear Kernel
2.) Polynomial Kernel
3.) Gaussian RBF kernel
4.) Sigmoid Kernel

5. EXPERIMENTS AND DISCUSSIONS

The effectiveness of the suggested technique has been tested using a number of methods, including: a) Global script recognition accuracy b) Accurate identification of words as script when their font sizes and styles are omitted from the training sample.

5.1. DATA SET PREPARATION

Because there were no standard databases available, authors created their own data set to demonstrate the effectiveness of the suggested features. The materials are scanned at a resolution of 300 dpi and printed on laser printers in both Gujarati and English. Hence, 5000 Gujarati words and 5000 English words were used to prepare the data set of 10,000 words.

5.2. GLOBAL SCRIPT RECOGNITION ACCURACY

Here, the 5-fold cross validation scheme is employed to assess the recognition outcome. First, 5-fold cross validation indices of each dataset's length are formed at random. The integers 1 through 5 are distributed equally among these indices. These numbers are used to specify the division of a large data set into five distinct sections. For each data set, the remaining four subsets are utilized for training, while one subset is used for testing. This is carried out five times for a classifier SVM with varied parameters, moving the testing data set to a different subset each time and taking the remaining subsets into consideration for training. Hence, a total of 5 sets of feature vectors that contain a 4:1 ratio of training and test data are obtained.

Several SVM classifier kernel functions are used in the studies. The fundamental reason for performance variations across various kernel SVM classifier functions is related to the distribution of feature data. Here, statistical features are tested using linear, polynomial, and RBF kernels. Figure 4 displays the SVM classifier's average accuracy for word recognition in Gujarati and English using various kernel functions.

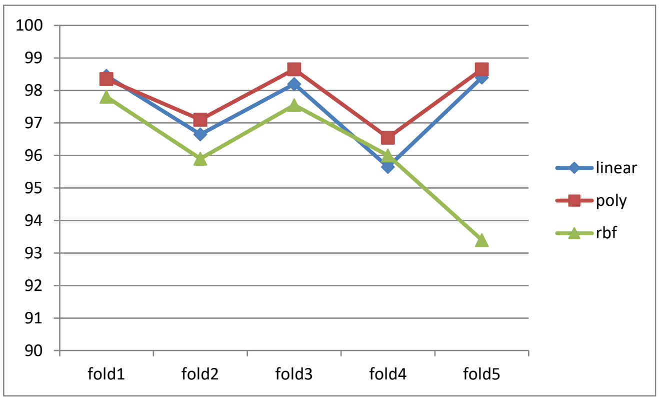


Fig. 4. Average Recognition Accuracy using different classifier

It should be highlighted that proposed features with linear and polynomial kernel functions had average accuracy values of 97.47% and 97.86%, respectively, outperforming RBF kernel function, which had a value of 96.13%. The confusion matrix for Gujarati and English words using the linear, polynomial, and RBF kernel functions of SVM is shown in Table 1.

Table.1. Confusion matrix of various classifiers

Classifier	Words	Gujarati	English
SVM-Linear	Gujarati	4827	173
	English	80	4920
SVM-Polynomial	Gujarati	4852	148
	English	56	4944
SVM-RBF	Gujarati	4753	247
	English	140	4860

6. CONCLUSION AND FUTURE ENHANCEMENT

In this study, classifier accuracy and word-level script identification are both attempted. In this work, structural traits and statistical features are both utilized. For the purpose of identifying the script of printed Gujarati and English words, SVM classifiers are utilized using various kernel functions. Results showed that the SIM-Polynomial kernel produced the greatest accuracy of 97.86%. It has never been done before to distinguish between Gujarati and English scripts at the word level. Only Gujarati and English words are used here to describe the work, but it can be expanded to include other Indian and non-Indian context in the future, and accuracy can be increased by taking other factors into account.

REFERENCES

[1]. Ghosh D., Dube T., Shivaprasad A. P., Script Recognition A Review. IEEE, Transactions on Patter Analysis and Machine Intelligence vol. 32, no. 12, pp. 2142-2161, 2010.

[2]. Chaudhari S., Gulati R., A Survey on Script Identification in Multi-script Indian Documents. VNSGU journal of Science and Technology Vol 3, Issue 2, pp. 138-152, 2012.

[3]. Chaudhuri.B.B, Pal.U, An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi). In Proc. 4th ICDAR, Uhn. 1997.

[4]. Pal U., Chaudhuri B.B., Script Line Separation from Indian Multi-Script Documents. Proc. Int'l Conf. Document Analysis and Recognition, pp. 406-409, 1999.

[5]. Pal U., Chaudhuri.B.B, Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line. Proc. 6th Intl. Conf: Document Analysis and Recognition (ICDAR'OI). pages 790-794, 2001.

[6]. Padma M.C., Vijaya P.A. Global Approach for Script Identification using Wavelet Packet Based Features. International Journal of Signal Processing, Image Processing and Pattern Recognition. Vol. 3, No. 3, 2010.

[7]. Patil B., Subbareddy N.V. Neural network based system for script identification in Indian documents. Sadhana Vol. 27, part-i1, pp 83-97, 2002.

[8]. Dhandra B.V., Nagabhushan P., Hangarge M., Hegadi R., Malemath V.S., Script Identification Based on Morphological Reconstruction in Document Images. Proc. IEEE Int'l Conf. Pattern Recognition. vol. 2, pp. 950-953, 2006.

[9]. Vikram T.N., Guru D.S. Appearance based models in document script identification. ICDAR '07 Proceedings of the Ninth International Conference on Document Analysis and Recognition. Volume 02, 2007.

[10]. Dhanya.D, Ramakrishnan.A.G, Peeta B. P. Script Identification In Printed Bilingual Documents. Sadhana, Vol. 27, Part-1, Pp. 73-82, 2002.

[11]. Sukalpa C., Pal S., Katrin F., Pal U. Two-stage Approach for Word-wise Script Identification. 10th International Conference on Document Analysis and Recognition. 2009.

[12]. Pal U., Sinha S., Chaudhuri B.B. Multi-Script Line Identification from Indian Documents. Proc. Int'l Conf. Document Analysis and Recognition. pp. 880-884, 2003.

[13]. Kunte R.S., Sudhaker S. A Bilingual Machine-Interface OCR for Printed Kannada and English Text Employing Wavelet Features. 10th International Conference on Information Technology. 2007.

[14]. Aparna KG, Dhanya D., Ramakrishnan AG, Bilingual (Tamil – Roman) Text Recognition on Windows, Tamil Internet. California, USA 2002.

[15]. Dhandra BV, Mallikarjun H., Hegadi R., Malemath VS Word–wise Script Identification based on Morphological Reconstruction in Printed Bilingual Documents. In the proc. of IET International Conference on Vision Information Engineering VIE, Bangalore pp. 389-393, 2006.

[16]. Dhandra BV, Mallikarjun H. On Separation of English Numerals from Multilingual Document Images, In the journal of multimedia Vol 2, No 6, pp. 26-33, 2007.

[17]. Cortes C, Vapnik VSupport vector network. Machine Learning. , 20:273–297, 1995.

[18]. S. Chaudhari and R. Gulati, Script Identification from bilingual Gujarati-English Documents, International Journal of Computer Applications (IJCA), Vol. 9, 2014.

[19]. Dhandra B. V. et al., Script Identification of Camera Based Bilingual Document Images Using SFTA Features, International Journal of Technology and Human Interaction Volume 15, Issue 4, October-December 2019

[20]. Mahajan S. and Rani R., Word Level Script Identification Using Convolutional Neural Network Enhancement for Scenic Images, ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 21, No. 4, Article 83., 2022.