International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 6 ǁ June 2023 ǁ PP. 207-213

A Simple Approach for Language Identification from Trilingual Handwritten Documents using Discriminating Features

*1Aishwarya M Y, 2Anvika, 3Dr. Mahesh K Kaluti

1 VI semester Student, Department of Information Science and Engineering, 2 VIII semester Student, Department of Computer Science and Engineering, 3 Assoc. Professor, Department of Information Science and Engineering, PES College of Engineering, Mandya, India

**Corresponding Author: 1Aishwarya M Y Email: aishwaryamy19@gmail.com*

Abstract

In a multilingual country like India, a document may contain text lines in more than one language forms. For Optical Character Recognition (OCR) of such a multilingual document, it is necessary to identify different language forms of the input document, before feeding the documents to the OCRs of individual language. In this paper, a simple but efficient technique of language identification for Kannada, Hindi and English text lines from a handwritten document is presented. The proposed technique is based on the discriminating features of individual text lines of the input document image. The discriminating features are extracted from the preprocessed text lines of all the three languages and stored in knowledge base for later use during decision making. For a new test line, required features are extracted and compared with the stored knowledge base. A k-nearest-neighbor classifier is used to identify the language type of the text line. The proposed algorithm is tested on 600 handwritten text lines and an overall classification accuracy of 95.83% is achieved. Experimental results demonstrate that relatively simple technique can reach a high accuracy level for identifying the text lines of Kannada, Hindi and English languages.

Keywords: Tri-lingual document, discriminating features, knowledge base, KNN classifier.

Date of Submission: 02-06-2023 Date of acceptance: 13-06-2023

I. INTRODUCTION

In a multilingual environment automatic language identification is a challenging task while processing large volumes of handwritten document images. Also it is very essential to reliably identify the language type using the least amount of textual data when dealing with handwritten documents that contain multiple languages. Offline language identification serves as a forerunner for a multi lingual Optical Character Recognizer (OCR). OCR is a software used for digitizing the handwritten or printed document images. Most of the OCRs are designed for dealing with a single script/language. So it can not convert a document which contains more than one language. A country having multi lingual culture like India, multi lingual OCR is an essential software requirement for automation of handwritten documents processing. Automatic language identification techniques are useful in a wide variety of applications and a few of them are to sort document images of the same type; to select specific OCRs and to search online archives of document image for those containing a particular language.

In a multi-script multi-lingual country like India (India has 18 regional languages derived from 12 different scripts [1]), a handwritten document may contain text lines written in more than one script/language forms. Under the three language formulae [1], adopted by most of the Indian states, a handwritten document in a state may be written in its respective official regional language, the national language Hindi and also in English. Accordingly, handwritten documents seen in Karnataka, a state in India, may contain text lines written in its official regional language Kannada, national language Hindi and also in English. With this context, this paper aims at identifying language type from a trilingual handwritten document containing Kannada, Hindi and English languages.

The complexity of language identification becomes more challenging in the case of handwritten documents since the handwriting styles differ by individuals. Also the difference and the complexity of the text words increase due to the writing style, shape, and size of the text words. Thus handwritten documents present three main challenges in language identification [4]. For example, first, some languages resemble each other more when handwritten than when printed. Second, handwriting styles are more diverse than printed fonts.

Cultural differences, individual differences, and even differences in the way that people write at different times, increase the possible character and word shapes seen in handwritten documents. Third, problems typically addressed in preprocessing, such as ruling lines and character fragmentation due to low contrast, are more common in handwritten documents due to the variety of papers and writing instruments used [4].

Because of these challenging issues it is impossible to successfully apply the same template matching approaches that are used for machine printed documents [1, 5-7, 10] to identify languages from handwritten documents. Because of the greater variability of handwriting styles, sufficient numbers of reliable templates could not be identified. Hence, in this paper, a feature-based approach in which each document is characterized by a distinct set of features for each language is proposed.

1.1 Literature survey:

There has been extensive effort for the design of effective handwritten-language identification systems, which can be applied to variety of applications such as banking processes, the reading of postal codes, and so on. Trieu Son Tung [3] have given the survey about the feature extraction technique which is based on the overlap area between the characters, the run-length histogram, the denseness of the black pixels, and the width, height, and area in each word block. Trieu Son Tung [3] have highlighted that Kuhnke et al. have utilized the structural features of the characters, where straight lines that follow both the horizontal and vertical orientations and the symmetry between two sides are the key features. Judith Hochberg et.el., [4] proposed a classification method for which several structural word-block features including deviation of width, height, area, density of the black pixels are used. Hochberg, J [6] proposed an automatic script-identification process for the document images for which cluster-based templates are used. In this method, the texture based features are extracted from the text. Due to the variations of the textual appearance and also the writing style used by different users it is more challenging to apply the methods developed for one language type to apply for other handwritten text words [2]. Prasanthkumar et. el. [10], have made a survey on script and language identification for handwritten document images using different types of techniques along with the advantages and drawbacks of each techniques. It is seen that sufficient amount of work on language identification for printed documents have been carried out [10]. But very less amount of work on language identification for handwritten documents is reported in the literature. Hence, in this paper, a feature-based approach for identifying the language type from handwritten documents is proposed.

This paper is organized as follows. The Section 1.2 describes the proposed model, discriminating features and the algorithms used for language identification. The details of the data set used in the experiments are given in Section 1.3. The details of the experiments conducted and the results obtained are presented in Section 2. Conclusions are given in Section 3.

1.2 The Proposed Model:

It is observed that every language defines a finite set of text patterns and hence exhibits its own distinguishing features. The proposed model is inspired by a simple observation that every script/language defines a finite set of text patterns, each having a distinct visual appearance [5]. The character shape descriptors take into account any feature that appears to be distinct for the language [5] and hence every script could be identified based on its discriminating features. The texture of one particular language exhibits common features in the case of machine printed text lines. However, the texture of the same language exhibits different type of textures since different writers have different hand writing styles and hence texture based features cannot be used for identifying the language type of handwritten documents. Hence in this paper the proposed model is developed by extracting the distinct characteristic features at the micro level from each connected component of the text lines of three languages from the input document image.

1.2.1 Properties of the three languages – Kannada, Hindi and English

Languages/scripts are made up of different shaped patterns to produce different character sets. Individual text patterns of one language are collected together to form meaningful text information in the form of a text word, a text line or a paragraph. The collection of the text patterns of the one script exhibits distinct visual appearance and hence it is necessary to thoroughly study the discriminating features of each language that are strong enough to distinguish from other languages/scripts [10]. The most common distinct characteristic features of the three languages seen in handwritten documents irrespective of the writers and writing style are explained below.

Discriminating features of Kannada language:

Kannada text words consists of combination of vowels, consonants, modified consonant and/or compound characters. The compound characters have descendants called ‘vathaksharas’ found at the bottom portion of a character. The presence of these ‘vathaksharas’ could be used as a feature to identify the text word as a Kannada language. It could be observed that most of the Kannada characters have horizontal linear line structures present at the top portion of the characters. Also, it could be observed that majority of Kannada characters have upward curves present at their bottom portion. Thus, the presence of the structures like – horizontal linear lines, upward curves and descendants could be used as the supporting features to identify the text lines of Kannada language.

Discriminating features of Hindi language:

In Hindi, it is noted that many characters of these alphabets have a horizontal linear line at the upper part, which is called the headline. When two or more characters sit side by side to form a word, the headline portions touch one another and generate a long linear headline. Most of the pixels of these headlines form a linear line found at the top portion of Hindi language which can be treated as a strong discriminating feature of Hindi language. Another strong feature that could be noticed in Hindi language is the presence of vertical linear lines. These characteristic features are used to separate a Hindi text line.

Discriminating features of English language:

It is observed that the most of the English characters are symmetric and regular in the pixel distribution. One of the distinct and inherent characteristics of most of the English characters is the existence of vertical and slant line-like structures/strokes. The presence of vertical and slant line-like structures/strokes can be used as a supporting feature to identify an English text line. It could be observed that the upward-curve and downward curve shaped structures are present at the bottom and top portion of majority of English characters respectively. So, it was inspired to use such curve shaped structures also as the supporting features for identifying English language.

1.2.2 Feature Extraction

The input document image is preprocessed by removing noise and ruling lines which is very common in handwritten documents. From the preprocessed image, the distinct features of each language are extracted and the feature values obtained for each language are used in the proposed model. The basic concept used in the feature extraction technique is the eight-connected black component [8]. Using the 8-connectivity connected components the preprocessed document image is segmented into several text words. A bounding box is fixed for each text word using its leftmost black pixel, rightmost black pixel, topmost black pixel and bottommost black pixel [9]. Then, a set of six features are extracted from each bounding box. Majority votes are taken from a text line to classify the text line as a language type.

Terms used in the feature extraction technique:

By thoroughly observing the structural outline of the characters of the three languages, it is observed that the distinct features are present at a specific portion of the characters. So, the discriminating features are identified by partitioning the text word into three zones [9].

The row at which the first black pixel lies at the top of a word is called the top-line. The row at which the last black pixel lies in a word is called the bottom-line. The portion of the text word having the maximum number of black pixels in the middle portion of the text word is called the middle zone. The first row of the middle zone is called the mid-top-row and the last row of the middle zone is called the mid-bottom-row. Using the four lines – top-line, bottom-line, mid-top-row and mid-bottom-row as the reference lines, each word is divided into three zones – top-zone, mid-zone and bottom-zone. The term ‘top-zone’ is used to represent the portion of the text line between top-line and mid-top-row. Similarly, the term ‘mid-zone’ is used to represent the portion of the text line between mid-top-row and mid-bottom-row and the term ‘bottom-zone’ is used to represent the portion of the text line between mid-bottom-row and bottom-line. It is observed that the density of the pixels varies from one language to another in the three zones. Generally, the density of the black pixels in top zone and bottom zone is less and the density of the black pixels in the middle zone will be more. The distinct features are extracted from each zone of the text word of the respective language.

Different text words are partitioned in different ways, as different shaped characters are present in text words. Generally, a text word can be partitioned into three zones as explained above, only when the four reference lines namely – top-line, top-max-row, bottom-max-row and bottom-line are obtained. A partition with at least three pixels height (fixed through experimentation) is considered as a top-zone or bottom-zone. For some text words, where top-line and top-max-row occur at the same location, top-zone is not obtained. Similarly, for some text words, if bottom-max-row and bottom-line occur at the same location, and then bottom-zone is not obtained. Also, for some text words, when the text word without the descendant is partitioned, then the bottom-zone is not obtained. Similarly, a text word without ascendants does not possess top-zone. In this paper only the nature of English characters are explained. For example, some characters of English language such as ‘b, d, f, h, k, l, t’ have ascendants and some characters such as ‘g, j, p, q, y’ have descendants. So, if the characters of the text word have ascendants, then the top-zone is obtained and if the characters of the text word have descendants, then the bottom-zone is obtained. For other characters like 'a, c, e, m, n, o, r, s, u, v, w, x, z', there are no ascendants and descendants. For the text words having these characters, top-zone and also bottom-zone are not obtained. So, only middle-zone is obtained for such text words. Because of the distinct features exhibited by the three languages, corresponding discriminating features are extracted from the three languages as explained below.

Feature 1: bottom-component:

The presence of vathaksharas or descendants found at the bottom portion of Kannada language could be used as a feature called bottom-component. The feature named 'bottom-component' is extracted from the bottom portion of the input text line.

Through experimentation, it is estimated that the number of pixels of a descendant is greater than 8 pixels and hence the threshold value for a connected component is fixed as 8 pixels. Any connected component whose number of pixels is greater than 8 pixels is considered as the feature bottom-component.

Feature 2: Top-horizontal-linear-line:

It is observed that the horizontal line like structures is present at the top portion of the text lines of Kannada and Hindi languages. The connected components present at the top portions of the text line are analyzed. If the number of pixels of these connected components is greater than the 75% of the height of the text line then such components are used as the feature named top-horizontal-linear-line. The presence of this feature is analyzed using 600 text lines of each of the three languages Kannada, Hindi and English. From the experimental analysis, it is observed that the presence of the feature top-horizontal-linear-line is more in Kannada and Hindi language and it is almost absent in the case of English language. So, using the feature named top-horizontal-linear-line, the text lines of Kannada and Hindi languages could be separated from English language. The length of the feature top-horizontal-linear-line (length of the feature top-horizontal-linear-line is measured with the number of pixels present in it) is different for different languages and hence, distinct threshold value for each language is computed through experiments. If the length of the feature top-horizontal-linear-line is greater than two times the text line height, then Hindi text line can be separated from Kannada line. Thus, using the length of the feature - top-horizontal-linear-line, text lines of Hindi language can be well separated from Kannada language.

Feature 3: Vertical-slant-line:

It is noticed from the handwritten text words of Hindi and English languages that they have vertically slant line segments. These vertical lines are present in the middle-zone of the text word which needs to be extracted. By convolving a vertical line filter over the image of the middle-zone, vertical lines like components are extracted and these components are used as the feature named vertical-slant-line. The presence of this feature is more in Hindi and English language and is absent in Kannada language. Hence, the feature vertical-slant-line can be used to discriminate text words of Hindi and English language from Kannada language.

Feature 4: Curve-shapes:

Most of the characters of Kannada language have double upward curve shaped components at the location of their bottom-max-row. This double upward shaped component is used as distinct feature for Kannada language. The structural shape of the Kannada language is observed thoroughly and noticed that left curve and right curve shaped components are present at the middle portion of a text word.

By thoroughly observing the structural shape of the English language, it is observed that the downward shaped components are present at the region of top-max-row. Similarly, the upward shaped components are present at the region of bottom-max-row of most of the characters of Kannada and English languages.

Detecting the curve shaped portion from a character is the key for extracting the features used for identification. The presence of a curve is obtained by verifying the variation between the two pixels of a connected component that appear on the same scan line for the complete scan of the component. The increasing variations of the two pixels for the entire scan of the component results in the downward curve shaped segment and the decreasing variations of the two pixels for the entire scan of the component results in the upward curve shaped segment. Similarly, the presence of a left curve is obtained by verifying the distance between two pixels of a connected component that appear on the same vertical scan line of a component. The increasing variations of the two pixels for the entire vertical scan of the component results in the left curve shaped segments and decreasing variations of the two pixels for the entire vertical scan of the component results in the right curve shaped segments. Such curve shaped components with upward curve and downward curve shapes are shown in Figure 1(a) and (b) respectively. Similarly, curve shaped components with left curve shape and right curve shape in a character are shown in Figure 2(a) and (b) respectively.

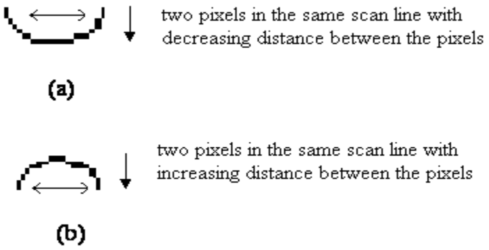

two pixels in the same scan line with decreasing distance between the pixels
(a)
two pixels in the same scan line with increasing distance between the pixels
(b)

Figure 1 Components showing
(a) Upward curve shape in a character
(b) Downward curve shape in a character.

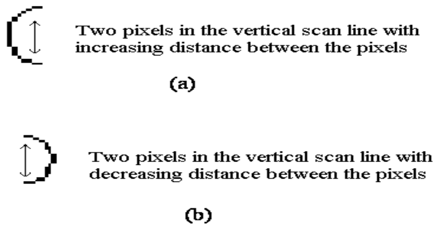

Two pixels in the vertical scan line with increasing distance between the pixels
(a)
Two pixels in the vertical scan line with decreasing distance between the pixels
(b)

Figure 2 Components showing
(a) Left curve shape in a character
(b) Right curve shape in a character.

Thus the four discriminating features of the three Kannada Hindi English used in the proposed model are shown in the Table 1. The letter ‘Y’ indicates that the corresponding feature is used for identifying the language and the letter ‘N’ indicates that the corresponding feature is not considered for that language.

Table 1: The discriminating features present in Kannada, Hindi and English language

Features used	Kannada	Hindi	English
Feature 1:bottom-component	Y	N	N
Feature 2: Top-horizontal-linear-line	Y	Y	N
Feature 3: Vertical-slant-line	N	Y	Y
Feature 4: Curve-shapes	Y	N	Y

1.2.3 Proposed Algorithm

The learning and classification algorithms of the proposed method are given below:

Algorithm 1:Learning ()

Input: Handwritten document image having Kannada, Hindi and English text lines.

Output: Knowledge base stored with the feature values of the three languages.

1. Preprocess the input document image.
2. Extract the four features from the preprocessed image.
3. Obtain features from a training data set 300 text images from each of the three languages and store these feature values in the knowledge base.

Algorithm 2: Classification ()

Input: Test document image in Kannada, Hindi and English languages.

Output: Script/language type of each text line of the test document.

1. Preprocess the test document image.
2. Obtain the features from the test document image as explained in the Algorithm 1.
3. Compute the distance between the feature values of the test document image with the feature values stored in the knowledge base using the Euclidian distance formula given below.

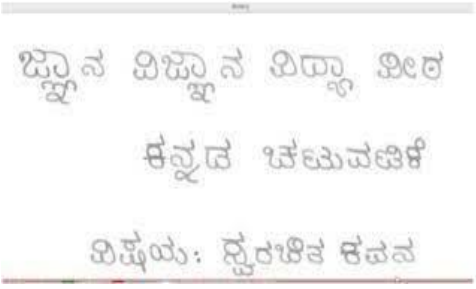
$$D(M) = \sqrt{\sum_{j=1}^{N}[f_j(x) - f_j(M)]^2}$$

where N is the number of features in the feature vector f, fj(x) represents the jth feature of the test sample ‘x’ and fj(M) represents the jth feature of Mth class in the feature library.

4. Classify the script/language type of the test sample ‘x’ using K-nearest neighbor classifier.

1.3 Dataset

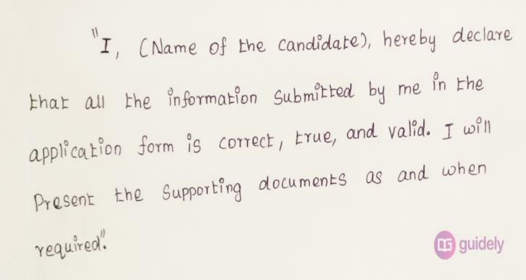
A dataset of handwritten documents containing 2000 text lines from the three languages - Kannada, Hindi and English is used for experimentation. 70% of the dataset is used for training and remaining 30% of the dataset is used for testing. The neatly handwritten documents are scanned and preprocessed to obtain images ready for further feature extraction technique. Input documents which contains neatly handwritten, well separated and also having linear text lines were considered for experimentation since the complexity increases to the worst level when text lines are overlapped and written non-linearly. Sample document image of Kannada, English and Hindi languages used in the proposed model are shown in Figure 1(a), 1(b) and 1(c) respectively. Some documents containing text lines of only one language and few documents containing text lines in combination of two and all three languages were considered for experimentation. As an initial work on language identification of handwritten documents, we have considered the dataset which contains neatly handwritten text words without having any overlapping text lines.



ಜ್ಞಾನ ವಿಜ್ಞಾನ ವಿದ್ಯಾ ಪೀಠ

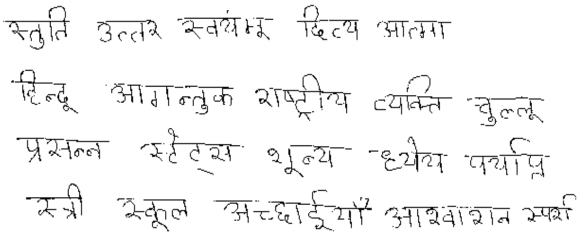
ಕನ್ನಡ ಚಟುವಟಿಕೆ

ವಿಷಯ: ಸ್ವರಚಿತ ಕವನ



"I, (Name of the candidate), hereby declare that all the information Submitted by me in the application form is correct, true, and valid. I will Present the Supporting documents as and when required."

Figure 1(a) Sample document image of Kannada   Figure 1(b) Sample document image of English



स्तुति उत्तर स्वयंभू दिव्य आत्मा

हिन्दू आगन्तुक राष्ट्रीय व्यक्ति चुल्लू

प्रसन्न स्टेट्स शून्य ध्येय पर्याप्त

स्त्री स्कूल अच्छाईयाँ आश्वासन स्पर्श

Figure 1(c) Sample document image of Hindi

II. RESULTS AND DISCUSSION

The proposed algorithm is trained using 1400 handwritten text lines and tested with 600 handwritten text lines of Kannada, Hindi and English languages. The test images are composed of handwritten text lines mixed with Kannada, Hindi and/or English languages. The test images containing text lines written in varied handwritten styles and sizes are also considered. But, the language type within a text line is assumed to be same. The percentage of classifying the three languages is given in Table 2. From the experiments on the test data set, the overall accuracy of the proposed algorithm has given 95.83%. From the experimental observations, it is noticed that the recognition rate falls down for the text lines with one or two text words. The proposed algorithm is implemented using MATLAB R2022a. The misclassification is when a text line consists of only one or two words and those words having one or two characters with minimal discriminating features.

Table 2. The percentage of classifying the three languages.

	Kannada	Hindi	English
Kannada (230 lines)	93.7%	2.5%	3.8%
Hindi (150 lines)	1.8%	98.2%	0%
English (220 lines)	4.4%	0%	95.6%

III. CONCLUSION

In this paper, an algorithm has been implemented for language identification of Kannada, Hindi and English text lines from handwritten documents. The simple approach is based on the analysis of the discriminating features extracted from the text lines of handwritten documents. Experimental results demonstrate that relatively simple technique can reach a high accuracy level for identifying the text lines of Kannada, Hindi and English languages. In the future work, dataset containing more complex handwriting and overlapped text lines will be addressed.

REFERENCES

[1]. M.C.Padma and P.A.Vijaya, "Identification of Kannada, Hindi and English Text Lines Using Profile Features", International Journal of Computer Science and Applications (IJCSA), ISSN: 0972-9038, © Technomathematics Research Foundation, Vol. 7, No. 4, pp. 16 - 33, 2010.

[2]. Wilensky, G., Crawford, T., Riley, R.: "Recognition and characterization of handwritten words". In Doermann, D. (ed.): Proceedings of the 1997 Symposium on Document Image Understanding Technology. College Park, MD: University of Maryland Institute for Ad- vanced Computer Studies 1997, pp. 87-98.

[3]. Trieu Son Tung and Gueesang Lee, : "Language Identification in Handwritten Words Using a Convolutional Neural Network", International Journal of Contents, Vol.13, No.3, Sep. 2017, https://doi.org/10.5392/IJoC.2017.13.3.038.

[4]. Judith Hochberg, Kevin Bowers, Michael Cannon, Patrick Kelly : "Script and Language Identification for Handwritten Document Images", International Journal on Document Analysis and Recognition (IJDAR) December 2000.

[5]. Spitz, A. L.: "Determination of the Script and Language Content of Document Images". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, no.3, 235-245 (1997)

[6]. Hochberg, J., Kelly, P., Thomas, T., Kerns, L.: "Automatic Script Identification from Document Images Using Cluster-based Templates". IEEE Transactions on Pattern Analysis and Machine Intelligence 19 : 2, 176-181 (1997)

[7]. Patent: Script identification from images using cluster-based templates (S-80,499)

[8]. Duda, T., Hart, P.: Pattern Classification and Scene Analysis. New York: John Wiley & Sons 1973, pp. 114-118

[9]. M.C.Padma and P.A.Vijaya, "Script Identification of Text Words from a TriLingual Document", International Journal of Image Processing (IJIP), ISSN (Online): 1985 – 2304, CSC Press, Computer Science Journals, Volume 4, Issue 1, pp. 35-52, March, 2010.

[10]. Prasanthkumar P V, Midhun T P, Archana Kurian: "A survey on Script and Language identification for Handwritten document images", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. V (Mar – Apr. 2015), PP 105-109, www.iosrjournals.org.