Word–Wise Script Identification from Indian Documents

Suranjit Sinha, Umapada Pal, and B.B. Chaudhuri

Computer Vision and Pattern Recognition Unit
Indian Statistical Unit
203 B.T. Road; Kolkata – 700 108; India
umapada@isical.ac.in

Abstract. In a country like India, a single text line of most of the official documents contains two different script words. Under two-language formula, the Indian documents are written in English and the state official language. For Optical Character Recognition (OCR) of such a document page, it is necessary to separate different script words before feeding them to the OCRs of individual scripts. In this paper a robust technique is proposed to extract word-wise script identification from Indian doublet form documents. Here, at first, the document is segmented into lines and then the lines are segmented into words. Using different topological and structural features (like number of loops, headline feature, water reservoir concept based features, profile features, etc.) individual script words are identified from the documents. The proposed scheme is tested on 24210 words of different doublets and we received more than 97% accuracy, on average.

Keywords: Script Identification, Indian script, Bangla script, Malayalam script, Gujarati script, Devnagari script, Telugu script, Multi-script OCR.

1 Introduction

In India there are 19 languages and 12 scripts are used for these languages. In a country like India, a single text line of most of the official documents contains two different script words. Under two-language formula, the Indian documents are written in English and the state official language. For Optical Character Recognition (OCR) of such a document page, it is necessary to separate different script words before feeding them to the OCRs of individual scripts [2]. In this paper a robust technique is proposed to extract word-wise script identification from Indian doublet form documents. Here, we consider five major Indian doublet documents. These doublets are {Devnagari, English}, {Bangla, English}, {Malayalam, English}, {Telugu, English}, and {Gujarati, English}.

Among the pieces of earlier work of different script line separation, Spitz [7] described a method for classification of individual text from a document containing English (Roman) and Japanese text. Later, Spitz [8] developed a method to separate Han based or Latin based script separation. He used optical density distribution of characters and frequently occurring word shape characteristics for the purpose. Using cluster based templates, an automatic script identification technique has been described by Hochberg *et al.* [4]. Wood *et al.* [10] described an approach using filtered pixel projection profiles for script separation. Ding *et al.* [3] proposed a method for separating two classes of scripts : European (comprising Roman and Cyrillic scripts)

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 310–321, 2004.

and Oriental (comprising Chinese, Japanese and Korean scripts). Recently, using fractal-based texture features, Tan [9] described an automatic method for identification of Chinese, English, Greek, Russian, Malayalam and Persian text. Previously, we have developed an automatic scheme for the text line identification from document containing Indian script [5]. Script characteristics, shape based features and statistical features are used for the purpose. All the above pieces of work deal with line-wise script identification.

In this paper a robust technique is proposed for word-wise script identification from Indian documents. In the proposed method, at first, the document is segmented into lines and then the lines are segmented into words. Using different topological and structural features (e.g. number of loops, headline feature, profile features, water reservoir concept based features like reservoir height and width, water flow direction, reservoir position, etc.) individual script words are identified from the documents.

The organization of the paper is as follows. In Section 2 properties of five major Indian scripts are discussed. We did not describe the properties of English since they are well known. Pre-processing like text digitization, noise cleaning, line and word segmentation etc. are described in Section 3. Different features used in the identification scheme are discussed in Section 4. Text word separation scheme is described in Section 5. Finally, experimental results and discussions are provided in Section 6.

2 Properties of Five Different Indian Scripts

Devnagari is the most popular script in India. It has 12 vowels and 33 consonants. They are called *basic characters*. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are written in this way they are known as *modifiers*. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as *compound characters* [1]. These type of basic characters, compound characters and modifiers are present not only in Devnagari but in other four scripts, except English. Bangla script has many similarities with Devnagari script. Modern Bangla script has 10 vowels and 32 consonant. In both Bangla and Devnagari alphabet it is noted that many characters have a horizontal line at the upper part, which is known as Shirorekha or Headline [1]. No English character has such characteristic and so it can be taken as a distinguishable feature to extract English from both Bangla and Devnagari. Malayalam script has 13 vowels and 37 consonants. Most of the characters in this particular script have a convex curve type shape at their left or right end or both. During the 2nd half of the 20th century a new written Telugu script evolved based on modern spoken language. Telugu script has 14 vowels and 34 consonants. Most of the characters in Telugu script has a '✓' like part in their upper region which is a distinct characteristic of this script. The Gujarati script was adopted from the Devnagari script. That's why this script has much similarity with Devnagari except that characters in Gujarati script has no Shirorekha type feature. Gujarati script has 11 vowels and 32 consonants. All the basic characters in these five scripts are shown in Fig. 1. A text word in these six scripts can be partitioned into three zones. The *upper zone* denotes the portion above the headline, the *middle zone* covers the portion of basic (and compound) characters below headline and the *lower zone* is the portion where some of the modifiers can reside. The imaginary line separating middle and lower zone is called *base line*.

Fig. 1. Basic characters of five Indian scripts.

3 Preprocessing

The images are digitized by a HP scanner at 300 DPI. The digitized images are in gray tone and we have used a histogram based thresholding approach to convert them into two-tone images (0 and 1). The digitized image may contain spurious noise pixels and irregularities on the boundary of the characters, leading to undesired effects on the system. Median filtering technique is applied here to reduce the noise. The lines of a text block are segmented by finding the valleys of the horizontal projection profile computed by row-wise sum of pixel values. The position where profile height is the least denotes one boundary line [1]. A text line can be found between two consecutive

boundary lines. Finding the valley of the vertical projection profile, a text line is segmented into words. If the width of a valley is greater than 2*W, we assume that the valley is the separator of two words. The computation of W is done in the following way. We compute horizontally the mode of the white runs occur between two consecutive black runs of a line. This mode is W and it generally represents the distance between two characters in a word. From the experiment we noticed that distance between two words is greater or equal to twice of the distance between two characters in a word. Hence, the threshold value for word segmentation is assumed as 2*W.

4 Features Used for Script Identification

The features are chosen with the following considerations (a) Presence in characters of some scripts and absence in characters of at least one script (b) Robustness, accuracy and simplicity of detection (c) Speed of computation and (d) Independence of fonts, size and style of the text. Some of the principal features used in the scheme are:

Shirorekha feature: If we take the longest horizontal run of black pixels on the rows of a text line then the length of such run in Bangla and Devnagari scripts will be much higher than that of English script. This is because characters in a word are connected by head-line in Bangla as well as in Devnagari script. For illustration, see Fig.2. Here row-wise maximum run is shown in the right part of the words. This run information has been used to separate Bangla and Devnagari scripts from English. But for other scripts this feature cannot work to separate English from those scripts. We say head-line feature exists in a word if one of the following two conditions satisfy in the word (a) if the length of the longest run is greater than 70% of the width of a word (b) if the length of the longest run is greater than 2 times of the height of middle zone.

Distribution of vertical stroke feature: One of the most distinctive and inherent characteristics of most of the English characters is the existence of vertical line-like stroke at the leftmost part or both left & right of the English character. This stroke in a character can be computed by measuring the vertical run of black pixel at the leftmost part of the character. The character is said to have vertical stroke if the length of a black pixels run is at least 70% of the character height. If in a particular word 40% of the characters satisfy this feature then the word is treated as English. If the text is written in italics style this vertical stroke detection method may not work. We use a profile based method for italics vertical line detection. We compute left/right profile of a character and we observe the behaviors of the profile in the region between the mean-line and base line. If all left/right profiles of this region have unique behavior (either all increasing or all decreasing mode) we decide that a vertical line exists in that character as shown in Fig.3. Here all the left profiles in the character are in decreasing mode from top to bottom. Hence, we assume that a italics vertical line at the leftmost part exists in the character.

Water reservoir principle based feature: The water reservoir principle is as follows. If water is poured from one side of a component, the cavity regions of the component where water will be stored are considered as reservoirs [6]. By top (bottom) reservoirs we mean the reservoirs obtained when water is poured from top (bottom) of the component. (A bottom reservoir of a component is visualized as top reservoir when water will be poured from top after rotating the component by 180°). Similarly,

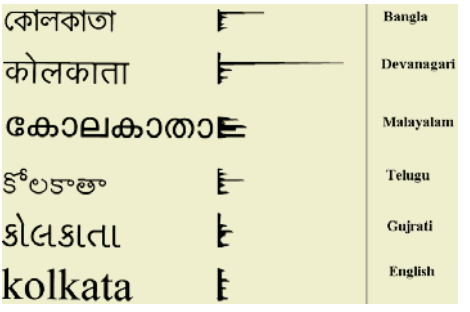

Fig. 2. Row-wise longest horizontal run is shown in five different script words.

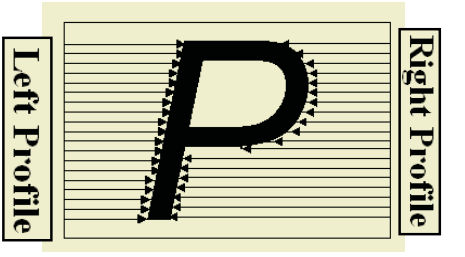

Fig. 3. Vertical line detection approach for italics character.

if water is poured from left (right) side of the component, the cavity regions of the component where water will be stored are considered as left (right) reservoirs. All these reservoirs are shown in Fig.4 for English character 'X'.

This water reservoir feature can be used as an useful tool to distinguish different script words. To identify English script word from other script words we use vertical stroke feature in most of the cases. Because this feature exists in most of the English characters. But there are some English words where many characters with vertical line may not occur. To identify such English words from other script words we use water reservoir concept based features. In English it is found that characters without vertical line have left/right or both left and right reservoirs. Now if a character has both left and right reservoir then the reservoir with the longest height is taken. If the total sum of this reservoir height and the stroke width of the character exceeds 70% of the character width then that character is marked. If for a particular word, 40% of the characters are marked then the word is identified as English. For illustration see Fig.5. In the word shown in Fig.5 has seven characters. Out of these seven characters four characters satisfy this reservoir property. Hence this word is treated as English. This feature is used to extract English from doublets like {Bangla, English}, {Devnagari, English}, {Malayalam, English}, {Gujarati, English}, {Telugu, English}.

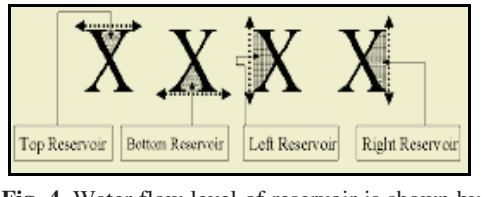

Fig. 4. Water flow level of reservoir is shown by dotted arrow.

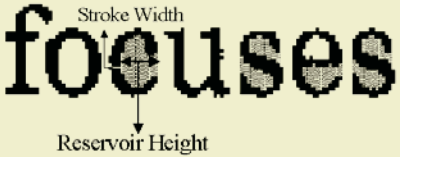

Fig. 5. Reservoir height and stroke width are shown.

The water reservoir feature is also used to identify Malayalam words from {Malayalam, English} doublet. To identify Malayalam words the water flow direction of reservoir is used. There are characters in Malayalam, which have top reservoirs where water flows from 'Right to Left' as shown in Fig. 6. Such characters occur frequently in Malayalam. But this kind of feature is absent in English characters. So if within a word at least one character has this feature then the word containing the character is

identified as Malayalam. The character shown in Fig. 6 is 'Pa' in Malayalam alphabet.

There are Malayalam characters which have top reservoirs whose height is almost that of the height of the character as well as there is a loop situated at the left portion of the reservoir. Example of a character, which has such property, is shown in Fig. 7. This is 'La' in Malayalam alphabet. The character with such property is totally absent in English. So if in a particular word at least one such character exists then that word is treated as Malayalam. These two reservoir based features are used to extract Malayalam words from English words.

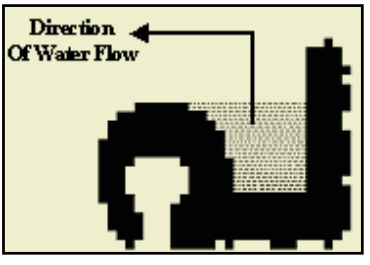

Fig. 6. Malayalam character contains top reservoir with water flow direction from 'Right to Left'.

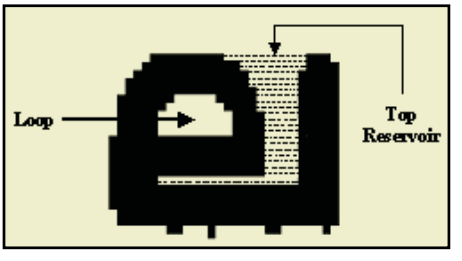

Fig. 7. Malayalam character contains loop within the left part of the top reservoir.

The reservoir based feature is also used to extract Gujarati from {Gujarati, English} doublet. There are Gujarati words with a single character as shown in Fig.8. This character is 'chha' in Gujarati alphabet. This type of character has a top reservoir whose height is greater than 80% of the character length. As this type of single character word is totally absent in English so the existence of such character identifies the Gujarati word from English.

The water reservoir concept based feature is also used to extract some Telugu words from English words. If water reservoir is considered from left side of a character then in Telugu script several characters can be found with left reservoir, whereas in English only five characters (a, s, x, y, z) have left reservoirs. Among them 's', 'x' and 'z' have both left and right reservoirs. So if left reservoir is found for a particular character in a word then that character is also tested for right reservoir. If the character has no right reservoir then that character is marked. In a particular word if the existence of such characters is 20% or more then the word is treated as Telugu.

Shift Feature Below Shirorekha

There are some Bangla/Devnagari words which may not satisfy the Shirorekha feature although some characters in the word may have head lines. For them, at first, each character of the word is segmented and leftmost black pixel of the head-line of each segmented character is noted. Let this pixel is x_1. Next head-line position of each segmented character is deleted, and after deletion of head-line, the first black pixel from top is noted. Let this pixel is x_2. If the distance between x_1 and x_2 is greater than or equal to 50% of the width of the character then that character is noted. If the number of such characters in a word is 20% or more then the word is treated as Bangla in case

of {Bangla, English} doublet and Devnagari in case of {Devnagari, English} doublet because this type of feature is almost absent in English. For illustration see Fig.9. which is a Bangla word. Though this word have Shirorekha on some character but as a whole it doesn't satisfy the Shirorekha feature but satisfy Shift feature below Shirorekha and treated as Bangla.

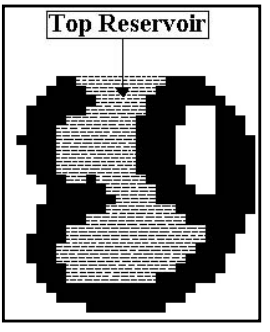

Fig. 8. Gujarati character with single Top Reservoir.

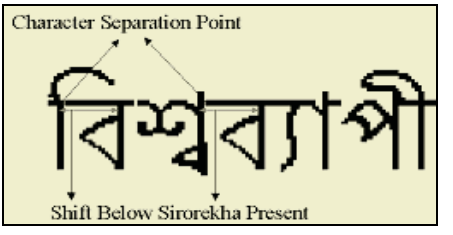

Fig. 9. Shift feature below Shirorekha Shown in a Bangla word.

Left and right profile: Suppose each character is located within a rectangular boundary, a frame. The horizontal or vertical distances from any one side of the frame to the character edge are a group of parallel lines which we call the *profile* (see Fig.10). If we compute left and right profile of the characters in a word, we can notice some distinct difference between Malayalam and English scripts. For example, in both the left and right profiles of Malayalam script, most of the characters have one transition point because of their convex shape and this feature is rare in English. By transition we mean change of the profiles from increasing mode to decreasing mode or vice-versa. We use left and right profile feature for the identification of Malayalam script. Left and right profile of a Malayalam character is shown in Fig.10.

There are many Telugu characters which have right profile with a single transition point, from decreasing to increasing. But in case of English this type of behavior is almost absent. So this profile based feature is used in {Telugu, English} doublet to extract Telugu characters from English. In a particular word if at least 20% of the characters satisfy this feature then the word is identified as Telugu.

Deviation feature: A character is first tested for right vertical line. If it has a right vertical line then the topmost and bottommost row value of that vertical line is noted. Then from the left bottommost part of the vertical line, anticlockwise rotation is performed up to 30% of the character length. During this traversal the leftmost column value is calculated for each row and the distance between two leftmost columns of two consecutive rows are noted. If one such difference is found which is greater than 1.5 times the stroke width among these differences then the row value corresponding to that difference is stored. If this row value is greater than 30% of the component length and the difference between this row value and the bottommost row value is greater than 1.5 times the stroke width then this character is marked provided that character has only one vertical line and that is a right vertical line. The definition of vertical line is already discussed earlier. If total number of such characters is greater

than 20% of the total characters in the word then the word is treated as Gujarati. This particular feature is almost absent in English and so this feature is used to extract English from Gujarati words. Fig.11 shows this particular feature.

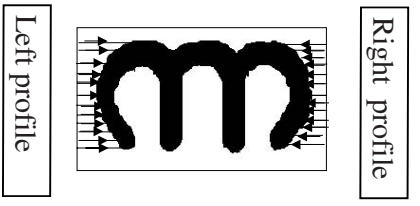

Fig. 10. Left and Right profile of a character is shown.

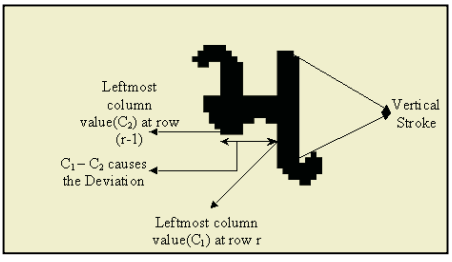

Fig. 11. Deviation feature of a character is shown.

Loop feature: Loop is defined as the hollow region enclosed by black pixels in a particular character. The size of this hollow region may vary from a single white pixel up to an area compatible with the area enclosed by the character itself. There are characters in some Indian scripts like Malayalam which have more than two loops of significant size. So, this feature can be used to distinguish English characters from Malayalam characters. Although, there are many characters in Malayalam with more than two loops, there exists no English character which has more than two loops. But if two consecutive characters touch then sometimes in English a components may have two loops. To overcome the connectivity problem it is first checked if the width of the character concerned is smaller than twice the length of the character height. If it is so and the character has more then two loops, then the word containing that character is considered as Malayalam in case of {Malayalam, English} doublet. One such Malayalam character is shown in Fig.12.

Tick feature: In most of the characters in Telugu script there is a distinct "tick" like feature called Telakattu at the top of a character. This feature is very helpful to identify Telugu script from English script because no such "tick" appears in the English characters. A Telugu character having a "tick" feature is shown in Fig.13. The position of "tick" feature in Telugu script is top of a character. To identify this "tick" feature in a character, at first, it is checked whether there is a top reservoir with left flow level at the upper part of the character or not. If two or more such reservoirs exist then the smallest and the uppermost one is to be considered. The water flow level of the reservoir is noted and let the water flow point be (X,Y). Starting from base point (x, y) of the reservoir the boundary up to the topmost point of the character is traced in a clockwise manner. Let the number of these traced points be N and these points be (x_i, y_i), i = 1….N. (By base point we mean the lowermost point of the reservoir). Similarly, from the base point of the reservoir anti-clockwise movement is performed along the boundary of the character up to the water flow point (X, Y). Let the number of these traced points be M and let these points be $(x_i^{/}, y_i^{/})$, i=1….M. For a character we say a "tick" feature exists if the following three conditions are satisfied. (a) N is greater than M, (b) $(y_1 - y) = (y_2 - y) = \ldots\ldots = (y_N - y)$ and (c) $(y - y^{/}_1) = (y - y^{/}_2) = \ldots\ldots\ldots = (y - y^{/}_M)$.

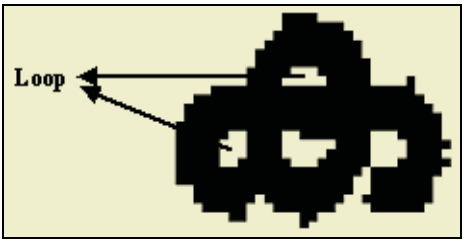

Fig. 12. Malayalam character containing more than two loops.

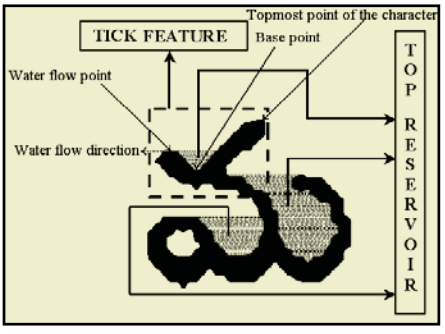

Fig. 13. Tick feature detection approach In a Telugu character.

Left inclination feature: There are some words with two or three characters in Bangla which do not satisfy the head-line feature or shift feature below Shirorekha but occur frequently in various sentences. For these words a new technique is proposed. It is found that most of these words contain some special type of characters which have a tendency of right to left inclination at the lower part of the character, as shown in Fig. 14. This Bangla character (named 'e') has this type of characteristic as shown in the diagram. This type of characteristic can be obtained as follows:

The lowermost part of each character is scanned left to right from bottom. The first pixel achieved in this way is stored. Starting from this pixel a clockwise rotation is performed until half of the character length is achieved. During rotation all the visited pixels are stored. If these pixels are all in decreasing manner and the length of this set of pixels i.e. the height of this inclined portion is greater than 60% of the half of the component length then it is said that a left inclination exists in the bottom portion of the character. If in a particular word one or more such characters satisfy this feature then the word is treated as Bangla. This feature is shown in Bangla character 'e' in Fig.14.

In Fig. 14 a Bangla word is shown which has only two characters. This particular word does not satisfy the Head-line or Shirorekha feature as well as Shift feature below Shirorekha. For these type of words left inclination feature is used. Here the Bangla character 'e' satisfies left Inclination feature as shown in Fig.14 and hence the word is treated as Bangla.

5 Script Identification Techniques

To extract English words from different Indian scripts a tree classifier is designed. In this classifier user first gives a choice (input) for a particular doublet, manually. Based on this input the control moves to the subsequent subtree corresponding to that doublet for classification. Here the subtree corresponding to each doublet is discussed separately, because of page limitation of this workshop, here we briefly present the identification scheme.

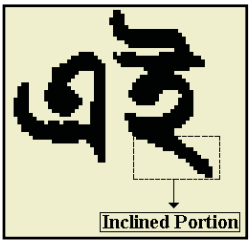

Fig. 14. Left Inclined portion is shown in a Bangla Character.

{Bangla, English} Doublet: The primary feature used in the subtree for this doublet is "Shirorekha" feature, because it is noted that the probability of occurrence of Headline or Shirorekha in a Bangla word is 0.997%. So it is justified to use "Shirorekha" at the top of the tree classifier. If this feature satisfies the word is identified as Bangla word. Otherwise the word is checked through shift feature. If shift feature satisfies the word is treated as Bangla, if not the word is tested by left inclination feature. If satisfies it's marked as Bangla otherwise tests for English words begin upon the word. First test is based on vertical stroke feature and the second one is based on water reservoir concept based feature. If the word satisfies any one of these two tests it is identified as English otherwise it is left as confused.

{Devnagari, English} Doublet: Similar subtree, which is used for {Bangla, English} doublet, is also used for {Devnagari, English} doublet due to the similarity in script characteristics between Bangla and Devnagari scripts.

{Malayalam, English} Doublet: In the subtree corresponding to this doublet convexity or profile based feature is taken as the primary feature because it is observed that most of the Malayalam characters have either left or right or both left and right profile as discussed earlier. So it is justified to take this feature as an initial feature. If the word satisfies the profile based feature then it is identified as Malayalam word, otherwise it is checked for reservoir based features as discussed earlier. If reservoir feature satisfies it is treated as Malayalam, otherwise it is checked whether any character of it has more than two loops or not. If such feature exists then it is Malayalam. Otherwise, vertical line feature is tested on the word. If vertical line like feature exists then the word is English. Else, water reservoir based feature is tested on it and if this feature exists the word is English. Otherwise, the word is left as confused.

{Gujarati, English} Doublet: As most of the Gujarati characters satisfy the deviation feature so it is taken as the principle feature to distinguish between English and Gujarati words in the subtree. Words satisfy the deviation feature are identified as Gujarati words. But the words not satisfying deviation feature are tested for a special feature known as upper zone feature. It is observed that most of the Gujarati words have a modified character at the upper part of the word. The length of the modified character is greater than any of the dots, which occur frequently in the upper part of English words. So if a modified character exists in the upper part of a given word whose length is greater than half of the component length then that word is treated as Gujarati. The words not satisfying this feature are checked by their size. If it is found that the word consists of a single character then water reservoir concept based feature for Gujarati character is applied to it. If this feature satisfies the word is treated as Guja-

rati. Otherwise the tests for English characters are applied. Vertical line feature and water reservoir concept based features are used for the purpose. If the word satisfies any of these features then it is treated as English, otherwise it is treated as confused.

{Telugu, English} Doublet: As most of the Telugu characters have the 'tick' feature so it is justified to use this feature as the principle feature in the subtree. If this feature is satisfied for a word then it is identified as Telugu, otherwise left reservoir feature is used for identification. If left reservoir feature satisfies, it is identified as Telugu. Otherwise, right profile based feature is used. If this feature exists then the word is identified as Telugu. If not then test for English word begins. If the word satisfies any of the two features (one is vertical line based feature and the second one is water reservoir based feature) then it is treated as English, otherwise it is left as confused.

6 Result and Discussion

For experiment two data sets were considered from different documents like question papers, bank account opening application form, money order form, computer printouts, translation books, dictionary etc. Some of the data captured from dictionary and money order application form were inferior quality. A data set of 7500 data was used for training and the other data set of 24210 data was used for testing of the propose schemes. We noted that the accuracy rates of script word separation schemes of five doublets {Devnagari, English}, {Bangla, English}, {Malayalam, English}, {Gujrathi, English} and {Telugu, English} were 97.14%, 98.30%, 98.89%, 97.22 and 98.06%, respectively. These statistics are computed from 5196, 5244, 5189, 5073 and 3508 script words of {Devnagari, English}, {Bangla, English}, {Malayalam, English}, {Gujrathi, English} and {Telugu, English} doublet documents. The overall accuracy of the proposed system was about 97.92%.

From the experiment we obtained highest accuracy in {Malayalam, English} doublet. This is because most of the Malayalam characters have a profile based feature and vertical line like feature does not exist in the side of the characters. Also, from the experiment we noticed that most of the errors are generated from poor and noisy documents which are mostly broken after digitization.

The confusion rates of the five doublets scripts are 1.92%, 3.69%, 1.25%, 5.03%, 4.36%, respectively. Also, we observed that most of the error comes from small words with number of characters three or less.

This scheme does not depend on the size of characters in the text line. Also we noticed that this approach is font and case insensitive. The use of simple features, which are easy to compute, make our system fast. Average execution time for an A4 size document on a P-IV machine is about 11 seconds.

The work presented here is a step towards building a multi-lingual OCR system that can work for all major Indian scripts. The next step is to build OCR modules for each individual scripts. To this end, successful modules for Bangla and Devnagari has already been demonstrated [2]. We encourage researchers to develop OCR technologies for other Indian scripts so that computer-based Indian document technology achieves its adulthood in near future.

References

1. B. B. Chaudhuri and U. Pal, A complete printed Bangla OCR system, *Pattern Recognition*, vol 31, pp 531-549, 1998.
2. B. B. Chaudhuri and U. Pal, An OCR system to read two Indian language scripts: Bangla and Devnagari*, In Proc. International Conference on Document Analysis and Recognition*, 18-20 August, 1997.
3. J. Ding, L. Lam and C. Y. Suen, Classification of oriental and European scripts by using characteristic features, *In Proc. 4th International Conference on Document Analysis and Recognition,* pp.1023-1027, 1997.
4. J. Hochberg, L. Kerns, P. Kelly and T. Thomas, Automatic script identification from images using cluster-based templates*, In Proc. 3rd International Conference on Document Analysis and Recognition*, pp 378-381, 1995.
5. U. Pal, S. Sinha and B. B. Chaudhuri, Multi-script line identification from Indian documents, *In Proc. 7th International Conference on Document Analysis and Recognition*, pp. 880-884, 2003.
6. U. Pal, A. Belaïd and Ch. Choisy. Touching numeral segmentation using water reservoir concept. *Pattern Recognition Letters*, 24:261-272, 2003.
7. A. Spitz, Multilingual Document Recognition, Electronic Publishing, Document Manipulation, and Typography, R. Furuta, ed. Cambridge Univ. Press, 1990, pp 193-206.
8. Spitz, Determination of the script and language content of document images, *IEEE Trans. on Patt. Anal. and Mach. Intelligence*, vol 19, pp 235-245, 1997.
9. T. N. Tan, Rotation invariant texture features and their use in automatic script identification, *IEEE Trans. on Patt. Anal. and Mach. Intelligence*, vol 20, pp 751-756, 1998.
10. S. Wood, X. Yao, K, Krishnamurthi and L. Dang, Language identification for printed text independent of segmentation, *In Proc. Int'l Conf. on Image Processing*, pp 428-431, 1995.