A Script Independent Technique for Extraction of Characters from Handwritten Word Images

Ram Sarkar
Dept. of CSE
Jadavpur University, Kolkata, India

Samir Malakar
Dept. of CSE
Jadavpur University, Kolkata, India

Nibaran Das
Dept. of CSE
Jadavpur University, Kolkata, India

Subhadip Basu
Dept. of CSE
Jadavpur University, Kolkata, India

Mita Nasipuri
Dept. of CSE
Jadavpur University, Kolkata, India

ABSTRACT

A script independent character segmentation from word images technique has been reported here. Word to character segmentation is an important preprocessing step of optical character recognition process. But in case of handwritten text, presence of touching characters decreases the accuracy of the technique of the segmentation of the characters from the word. In this paper, segmentation of handwritten word of four different scripts namely, Bangla, Devanagri, Gurmukhi and Syloti are considered as the test samples. All these scripts are characterized by the presence of a distinct line along the top of the most of the characters forming the words, called the headline or Matra. Unlike English script, the characters of these handwritten scripts and its components often encircle the main character, making the conventional segmentation methodologies inapplicable. For the segmentation technique two fuzzy features, to identify the *Matra* region and potential segmentation point, are used here. Experimental results, using the proposed segmentation technique, on sample of 400 handwritten word images containing all the above mentioned scripts of Bangla, Devanagri, Gurmukhi and Syloti show a success rate of 95.41%, 93.61%, 91.23% and 92.37% respectively.

Categories and Subject Descriptors

H. 2.0. **[Image]:** Image Processing, Computer Vision, Algorithm implementation and evaluation

General Terms

Algorithm, Performance, Experimentation

Keywords

Character segmentation, handwritten word images, Script independent technique, Fuzzy features.

1. INTRODUCTION

India is a multi-lingual country. There are many languages spoken in India. Most of them relate to one of the officially recognized languages. All these languages have a phonetic base, though their writing systems vary. Some of the languages have a common script and some have scripts of their own. There are 22 officially recognized languages in India. Out of these, Urdu, Sindhi and Kashmiri are primarily written in Perso-Arabic scripts, but get written in Devanagri too. Apart from Peraso-Arabic scripts, all the other 10 scripts used for Indian languages have evolved from ancient Brahmi script and have a common phonetic structure, making a common character set possible. The northern scripts are Devanagri, Gurmukhi, Bangla, and Gujarati etc. while southern scripts are Telegu, Tamil, and Kannada etc.

Devanagari 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. These types of basic characters, compound characters and modifiers are present not only in Devanagari but in other scripts, except English. The official language of India, Hindi is written in the Devanagri script. Devanagri is also used for writing Marathi and Sanskrit. It is also official script of Nepal.

Gurmukhi script is used in northwestern India, mainly in Punjab. In vocabulary it is very similar to Western Hindi. It has little literature and shows little borrowing from Persian, Arabic, or Sanskrit. Gurmukhi is devised by the Sikh Guru Angad (1539-52), in order to correct certain inadequacies so that sacred literature might be accurately recorded. Gurmukhi script alphabet consists of 41 consonants and 12 vowels. Besides these, some characters in the form of half characters are present in the feet of characters.

In the nineteenth and twentieth centuries a distinct script was widely used by the rural Muslim population of the Barak-Surma region in the northeastern part of the Indian subcontinent. Over the years the script came to be known as Sylhet Nagri or Syloti script [9] and formed the creative expression of the distinctive culture of the region. Syloti, an eastern Indo-Aryan language, is used by 10 million people in the three districts of Assam in India - Cachar, Hailakandi and Karimganj - and four districts of Sylhet in Bangladesh – Sadar Sylhet, Maulabi Bazar, Habiganj and Sunamganj. This is an ancient region the political cartography of which changed again and again, though cultural frontier remained porous. A continuous process of acculturation-accommodation marked the region and assimilation displayed a spirit of innovation. Syloti script is an expression of this creativity and the literature written in this script embodies the cultural history of the region. Is has 5 independent vowels and 5 dependent vowels attached with consonants and 27 consonants. Syloti is closely related to Bangla and most speakers are bilingual in Syloti and Bangla.

Bengali or Bangla is an Indo-Aryan language of the eastern Indian subcontinent, evolved from the Sanskrit language. Bangla is native to the region of eastern South Asia known as Bengal, which comprises present day Bangladesh, the Indian state of West Bengal, southern Assam- also known as Barak Valley, and part of Tripura. With nearly 230 million total speakers, Bangla is one of the most spoken languages (ranking 5th) in the world. Bangla is the national and official language of Bangladesh. It is the official language of the states of West Bengal and Tripura. It is also a major language in the Indian union territory of Andaman and Nicobar Islands. The Bangla script, with a few small modifications, is also used for writing Assamese. Other related languages in the region also make use of the Bangla alphabet. Meitei, a Sino-Tibetan language used in the Indian state of Manipur, has been written in the Bangla script for centuries, though Meitei Mayek has been promoted in recent times. The Bangla script has been adopted for writing the Sylheti language as well, replacing the use of the old Syloti script. Modern Bangla script has 11 vowels and 39 consonant.

In all of the four scripts mentioned above, it is noted that many characters have a horizontal line at the upper part, which is known as *Matra* or headline. No English character has such characteristic and so it can be taken as a distinguishable feature to extract English from these scripts. Most characters in these scripts have horizontal lines at the top, called the *headline* or *Matra*. In continuous handwriting, from left to right direction, the *Matra* of one character joins with the *Matra* of the previous or next character of the same word. In this fashion, multiple characters and Modified shapes in a word appear as a single connected component joined through the common *Matra*. It is also observed that all the characters and Modified shapes in a word appear to hang from the hypothetical *Matra* of the word, i.e., the *Matra* of each character in a word makes a uniform angle with the horizontal axis.

Optical Character Recognition (OCR) of text documents requires *Segmentation* of word images prior to *recognition.* Isolating individual alphabetic characters in the script image is often significant enough to make a decisive contribution towards the success rate of the overall system. Success of an OCR system for text documents highly depends on proper segmentation because each word segment produced in this process is a candidate character prior to recognition. Obviously, the more is the accuracy of segmentation, the less will be the error in recognition.

Due to infinite variability of handwritten characters, it is very challenging to segment the handwritten word images accurately.

Word segmentation is one of the core problems of OCR of handwritten text, which has long been an active area of research. Some important contributions so far made in this field include of English texts [1], [2].

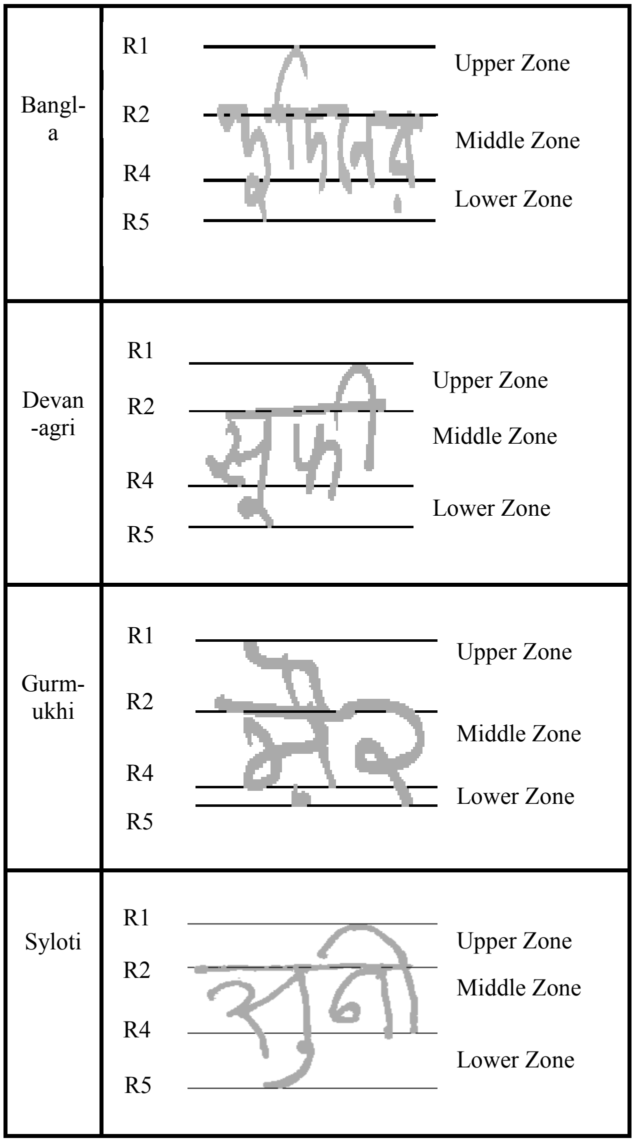

Figure 1: Illustration of some important features of all the four different scripts

The work relating to OCR of Bangla script, as mentioned earlier, is found to have few references in the literature. Two such instances [3], [4], one focusing on recognition of isolated handwritten characters based on stroke features and the other concentrated on a multistage approach based on different topological features respectively, have not addressed the problem of Bangla text segmentation. The problem of Bangla text segmentation has been addressed in [5]. The work has produced a complete OCR system for printed Bangla text. References involving segmentation of handwritten Bangla words are Recursive contour following [6] and water reservoir principle [7], separately applied in the past for this purpose.

Some important contributions of Gurmukhi script have been reported in [10], [11]. Segmentation of touching characters in Gurmukhi script is reported in [10] and the proposed technique in [11], segments the words in an iterative manner by focusing on presence of headline, aspect ratio of characters and vertical and horizontal projection profiles.

The work relating to OCR of Devanagri script is found in [12], [13], [14]. The problem of Devanagri text segmentation of printed document has been addressed in [13] and segmentation of handwritten document has been addressed in [14].

In our earlier work [8], we had applied our segmentation technique on Bangla script only. In the present work, we have extended our work of segmentation methodology by incorporating two fuzzy features and applied it on other scripts, namely, Devanagri, Gurmukhi, and Syloti also. As our technique gave very encouraging result on *Matra* based script, especially Bangla, we have attempted here to apply the technique to other above-mentioned *Matra* based scripts.

Any word, written in these scripts, can be partitioned horizontally into three adjacent zones. Figure 1 shows the three different zones of all the four scripts. The portion of each word on and above the *Matra* is identified as the 'upper zone'. The main body of the characters in a word and the portion of the word below the main body are identified as the 'middle zone' and the 'lower zone' respectively.

2. PRESENT WORK

The current methodology of the segmentation technique can be subdivided into the following steps. *Zone analysis* module identifies the 'upper', 'middle' and 'lower zones' in a word image. The *Matra identification* module identifies the probable *Matra* region at the top of the middle zone of the word. Different features, based on column wise pixel counts, are computed in the middle zone to identify potential segmentation points on the *Matra*. *Matra* pixels are hypothetically erased along the segmentation points and isolated segments are identified using 8-neighbour connected component labeling algorithm. Because of over segmentation, some characters get internally segmented in the middle zone. These broken components are recombined with the best possible main character segment. Due to the inherent complexity of handwritten words, the segmentation and recombination technique generate different types of segments in all the three zones of the word. The overall technique is illustrated in Figure 2.

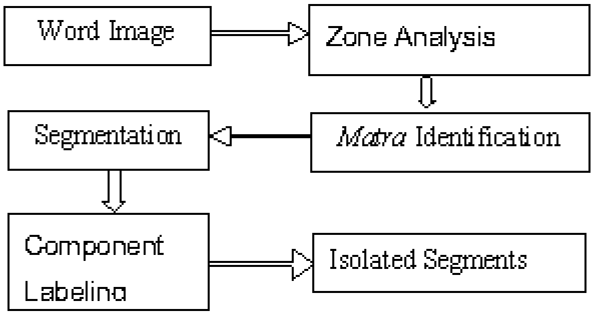

Figure 2: Schematic diagram of the segmentation process

3. IMPLEMENTATION DETAIL

The details of the present working methodology are discussed in the following subsections.

3.1 Analysis of 'Zones' from a Word Image

Characters and Modified shapes in a word often extend above the common *Matra* or appear below the main character body. In the current work, we have identified three adjacent horizontally partitioned zones from each word image. More specifically, the top row of the upper zone (R_1), the top row of the middle zone (R_2), the bottom row of the middle zone (R_4) and the bottom row of the lower zone (R_5) are identified from the word image. A horizontal pixel scan of the word image from top towards bottom identifies the first row with any black pixel as the top row of the upper zone. Similarly, a horizontal scan from bottom towards top identifies the first row with any black pixel as the bottom row of the lower zone. Identification of the top and bottom row boundaries of the middle zone is a challenging task in handwritten word segmentation.

In any word image, the common *Matra* appears as a boundary between the upper zone and the middle zone. Identification of the common *Matra* of a handwritten word is a difficult task. Ideally, a *Matra* of a word is a horizontal stripe of black pixels appearing at the top of the middle zone, touching most of the characters and Modified shapes in a word.

In the present work, to identify the common *Matra* of the word, horizontalness of each row is computed. Each black pixel of the word image is replaced by the length of the *longest run* of black pixels in horizontal direction through itself. Sum of the horizontal longest run values of all the pixels in a row is computed for each row of the word image. The row with the highest sum represents the row with maximum horizontalness. This row signifies the upper boundary of the middle zone.

To identify the lower boundary of the middle zone, horizontal transition points between text and background pixels are computed. In each row, starting from the top of middle zone to the bottom of the lower zone, the sum of transition points between text pixel to background pixel and vice versa are computed. The average number of row wise transition points in the middle and lower zone is computed as η. Computing the row wise sum of transition points from the bottom of the lower zone towards top, the first row with sum greater than η is identified as the lower boundary of the middle zone. The middle row of the middle zone is taken as R_3.

3.2. Fuzzy Matra Estimation

The common headline or *Matra* of a connected word segment may be identified as the continuous horizontal stripe of black pixels appearing at the top of most of the characters and some of modified shapes in the word segment. In a cursive handwriting the appearance of a *Matra* is often disjoint and wavy. This makes the identification of potential *Matra* pixels a challenging task. In the present work, we have developed two fuzzy measures to identify the membership value of each pixel, in the region R_1 to R_4, for its potential of belongingness to *Matra*.

3.2.1 Horizontalness feature

One of the important features of these scripts is the horizontalness property. This horizontalness property of the *Matra* may be extracted from the row wise sum of continuous run of black pixels, as shown in Figure 3. This value is normalized with respect to the maximum longest run value of any pixel within the word image. The row, which has the maximum longest run value, is taken as the central row of the *Matra* region. This row, which is the 1st approximation of the R_2, is called R_{21}.

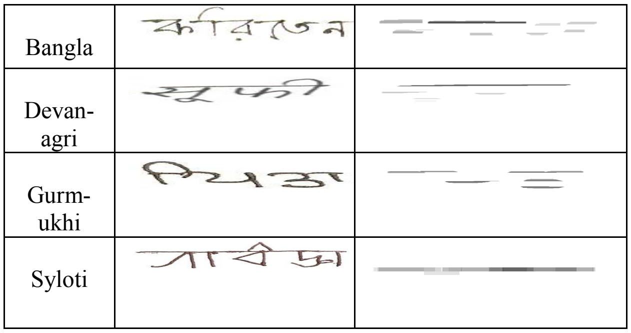
Bangla
Devan-agri
Gurm-ukhi
Syloti

Figure 3: Illustration of the horizontalness feature

3.2.2 Verticalness feature

Many characters and modified shapes in these scripts have vertical stripe of black pixels, as a part of their shapes. This vertical stripe often appears at the right side, middle or left side of the characters. These stripes touch the *Matra* of a word image and often extend till the bottom of the respective characters or modified shapes. In the present work, we have developed a technique to identify prominent vertical stripes in word image and identify their average top and bottom rows within the principal segments. This verticalness property of the *Matra* may be extracted from the column wise count of continuous run of black pixels, as shown in Figure 4. This value is normalized with respect to the maximum vertical longest run value of any pixel within the word image. We have considered those stripes which have the vertical longest run values greater than a threshold value. This threshold is chosen as the mean of all the longest run value in each column. Considering all the row numbers of top of all the selected vertical stripes, we have determined the average of all the values. We have taken this value as the 2nd approximation of the R_2 and called this R_{22}.

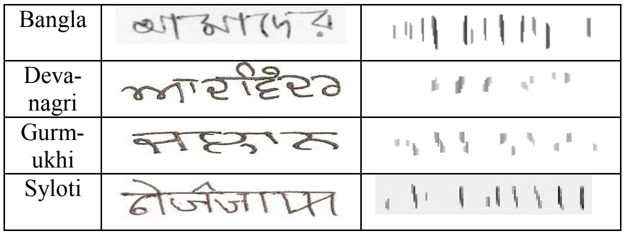
Bangla
Deva-nagri
Gurm-ukhi
Syloti

Figure 4: Illustration of the verticalness feature

Finally we have calculated the average of these two feature values i.e. R_{21} and R_{22}. This average value is then chosen as the final value of the central row of the *Matra* region i.e. R_2.

3.3 Design of the fuzzy membership function

In the present work, we have designed a bell shaped membership functions to map the horizontalness feature values of each row to determine its belongingness in the *Matra* region. The generalized bell function depends on three parameters a, b, and c as given by:

$$f(x;a,b,c)=\frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}}$$

Where, the parameter b is usually positive. The parameter c locates the center of the curve, i.e., R_2 and x is the row index for any black pixel $P_{x\,y}$ in the word image. For computation of the fuzzy feature values, we have designed a fuzzy function, viz, $f_h(x_h\ ,f(x;a,b,c))$ for horizontalness feature respectively. Such that,

$$f_h(x_h\ , f(x;a,b,c)) = x_h\ * f(x;a,b,c)$$

Where, x_h is normalized horizontalness component of each pixel $P_{x\ y}$ under consideration and $0 \leq x_h \leq 1$. . Figure 5 shows a diagramatic representation of the bell shaped fuzzy membership function, designed for the present work.

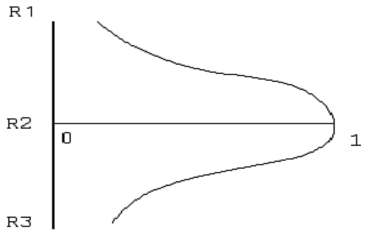

Figure 5. Fuzzy Bell-shape memberships function for Matra determination

A pixel $P_{x\,y}$ is identified as a headline pixel, if its value exceeds the mean of all such $f_h(P_{x\,y})$ values within the region R_1-$R_{3.}$

3.4 Design of fuzzy segmentation features

Once the black pixels constituting the *Matra* of a word segment are identified the next task becomes to identify certain column positions on the *Matra* from where the word segment can be vertically segmented into constituent characters. Such column positions are called terminal points of segments. One of the prominent features for identifying terminal points of segments is the number of black pixels along each vertical column position on the *Matra*. The less is the number of black pixels along a vertical column position on the *Matra*, the higher is its degree of belongingness (μ_1) to the set of terminal segment-points. On this basis a bell-shaped fuzzy membership function (μ_1), as discussed in previous section, is designed.

Another feature (F_2) is considered here within the region (R_2 – R_3). Here again the more is the distance, the less is the degree of belongingness (μ_2) of the associated point to the set of segment terminal points. The necessary membership functions (μ_1 and μ_2) for these features are shown in Figure 6.

To determine finally whether a black pixel on the *Matra* can be considered as a segment terminal point, the average of the two feature values exceed certain predetermined threshold, are finally considered as segment terminal points. The threshold is fixed up by taking the average of the two feature values of all the black pixel positions over the *Matra* of a word segment.

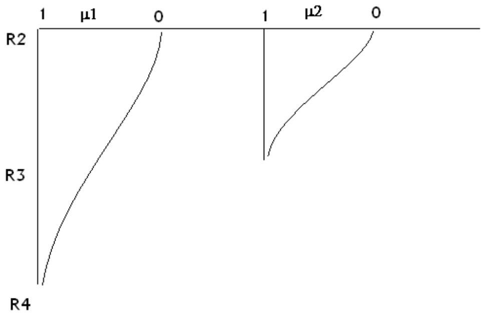

Figure. 6: Fuzzy membership functions µ1 and µ2

4. RESULTS AND DISCUSSION

In the present work, we have collected isolated handwritten Bangla, Devanagri, Gurmukhi, and Syloti word images from different persons of varying age groups. Word images are assumed to be slant and slope corrected and written in black ink with uniform pressure. Each such image is digitized using a flatbed scanner with 300 dpi resolution. 400 such word images (including all 4 scripts) were randomly selected for the current experimentation.

To evaluate the segmentation performance of the present technique the following expression is developed.

$$\text{Success rate} = (Ct / (Ct + Cu))*100$$

Where. Ct = the number of segment terminal points producing true segmentation, and Cu = the number of segment terminal points producing under segmentation. Whether a segment terminal point, identified by the present technique, produces true segmentation, under segmentation or over segmentation is determined through visual observation here.

The maximum success rate of the different scripts as achieved through the segmentation technique results, are 95.41%, 93.61%, 91.23% and 92.37% for Bangla, Devanagri, Gurmukhi and Syloti respectively. Table 1 shows the details of the results achieved by the technique.

Table 1: Details of the performance of segmentation technique on the different scripts

Script	Number of words	Number of true segment point	Under segmentation	Success rate (in %)
Bangla	150	603	29	95.41
Devanagri	100	425	29	93.61
Gurmukhi	100	385	37	91.23
Syloti	50	206	17	92.37

Figure 7 shows the images of some test samples of all 4 different scripts successfully segmented through this technique. Figure 8 shows some of the images where our technique fails to segments in the desired positions. The word segmentation technique, presented here for segmentation of handwritten scripts, can be considered as an important step towards the realization of a full-fledged OCR system of different scripts of Indian languages.

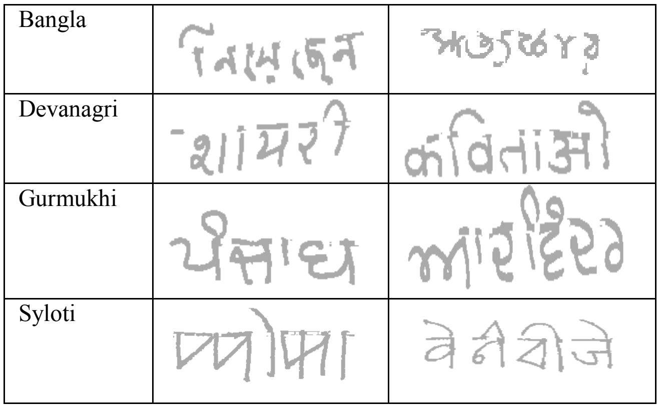

Figure 7: Sample word Images of different scripts showing the successful segmentation

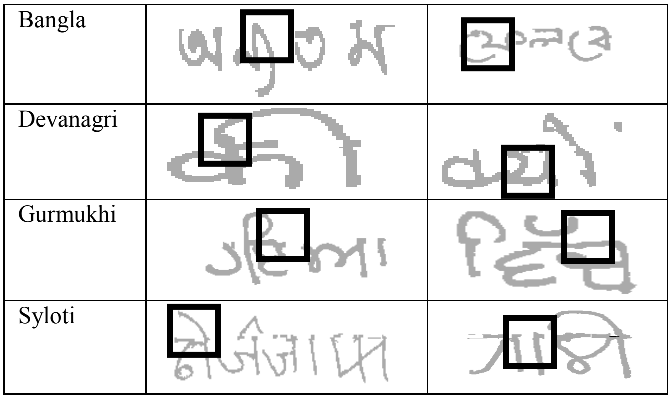

Figure 8: Sample word images of different scripts showing the failure cases (marked with the rectangular boxes) of the current technique

5. CONCLUSION

In this paper we have developed a methodology to segment the characters from the word images in four different scripts. All the different script used here, are having structural feature similar. More specifically, most of the characters of the scripts have a horizontal line at the top called *Matra*. To develop the technique we have estimated the *Matra* and the potential segmentation point using the fuzzy feature. This work is useful step towards the development of a multi-lingual OCR system.

6. ACKNOWLEDGEMENT

Authors are thankful to the “Center for Microprocessor Application for Training Education and Research”, “Project on Storage Retrieval and Understanding of Video for Multimedia” of Computer Science & Engineering Department, Jadavpur University, for providing infrastructure facilities during progress of the work.

7. REFERENCES

[1] R.G. Casey et.al. “A Survey of Methods and Strategies in Character Segmentation”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18,pp 690-706, 1996.

[2] R.M. Bozinovic et.al. “Off-line Cursive Script Word Recognition”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11,pp 68-83, 1989.

[3] A. F. R. Rahman, R. Rahman, M.C. Fairhurst, “Recognition of Handwritten Bengali Characters: a Novel Multistage Approach,” Pattern Recognition, vol. 35, p.p. 997-1006, 2002.

[4] T. K. Bhowmik, U. Bhattacharya and S. K. Parui, “Recognition of Bangla Handwritten Characters Using an MLP Classifier Based on Stroke Features,” in Proc. ICONIP, Kolkata, India, p.p. 814-819, 2004.

[5] A. Bishnu, B. B. Chaudhuri, “Segmentation of Bangla Handwritten Text into Characters by Recursive Contour Following,” in Proc. 5th ICDAR, pp. 402-405, 1999.

[6] U. Pal, S. Datta, “Segmentation of Bangla Unconstrained Handwritten text,” in Proc. 7th ICDAR, pp. 1128-1132, 2003.

[7] U. Garain, B. B. Chaudhuri, “Segmentation of touching characters in printed Devnagri and Bangla scripts using fuzzy multifactorial analysis,” IEEE Trans. On Systems, Man and Cybernetics – Part C: Applications and Reviews, vol. 22, pp. 449 – 459, 2002.

[8] S. Basu, R. Sarkar, N. Das, M. Kundu, M. Nasipuri, D. K. Basu, “A Fuzzy Technique for Segmentation of Handwritten Bangla Word Images”, International Conference on Computing: Theory and Applications (ICCTA), pp. 427-432, March-2007, Kolkata

[9] http://www.compcon-asso.in/projects/sylhet Nagri

[10] R. K. Sharma, A. Singh, “ Segmentation of Handwritten Text in Gurmukhi Script”, International Journal of Computer Science and Security, vol. 2, Issue 3.

[11] D. V. Sharma, G. S. Lehal, **“**An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script”, International Conference on Pattern Recognition – vol. 2, pp. 1022-125, 2006.

[12] R. M. K. Sinha, V. Bansal, “On Devanagari Document Processing”, IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada, 1995

[13] U. Garain, B. B. Chaudhuri, “Segmentation of touching characters in printed Devnagri and Bangla scripts using fuzzy multifactorial analysis,” IEEE Trans. On Systems, Man and Cybernetics – Part C: Applications and Reviews, vol. 22, pp. 449 – 459, 2002.

[14] V. Bansal, R.M.K. Sinha, “Segmentation of touching and fused Devanagari characters”, Pattern Recognition, vol. 35 (2002), number 4 pp. 875-893.