

Bengali Printed Character Recognition Using A Feature Based Chain Code Method

Ankita Sikdar¹, Sreeparna Banerjee^{2*}, Payal Roy¹, Somdeep Mukherjee¹, and Moumita Das¹

¹Department of Computer Science and Engineering, West Bengal University of Technology, Kolkata, West Bengal, India and ²Department of Natural Science, West Bengal University of Technology, Kolkata, West Bengal, India.

ankita.sikdar@gmail.com, sreeparnab@hotmail.com

ABSTRACT

Bengali, one of the official languages of the Indian subcontinent, is composed of 50 alphabets, of which 11 are vowels and 39 consonants. In addition, Bengali words are formed from compound characters and modifiers. Compound characters are formed by combining parts of single characters and modifiers are parts of vowels and consonants which make sense only when adjacent to or attached with a letter. In this paper, features of Bengali characters are studied using a hierarchical structure. The first few layers deal with features that broadly classify the characters into small size groups. The lower level features are more specific to each character within a group. Higher level features can be identified based on pixel density and arrangement, while the lower level features have been identified using a chain code technique. The algorithm progresses successively through each group in the hierarchy until it finds a match with the input character.

Keywords: Bengali character recognition, feature identification, chain code technique

1 INTRODUCTION

With the rapid proliferation of Internet and Mobile Computing in our daily lives, digitization of text for the purpose of storing and transmitting text across networks has become an absolute necessity. In order to perform digitization, Optical Character Recognition (OCR), both for handwritten and printed text, is required. OCR has thus become an active research area for document analysis and retrieval in different languages. Bengali is one of the official languages of India and the official language of Bangladesh. Hence, OCR in Bengali is also an important step in the digitization of Bengali printed characters processing applications also, the identification of the characters helps in knowing the text and using that information for further processing. A detailed description of its uses is described in [1].

A description of the research work carried out to identify printed characters and a discussion has been provided in [2]. In this paper we have used a chain code [3,4,5] based feature extraction method.

A comprehensive study for feature extraction has been presented in [6]. The feature extraction method that we present in this paper uses a hierarchical scheme. At first a feature set that introduces us to the features of the characters in a hierarchical manner is designed, with the top three levels being the basic features for all the characters and the later levels constitute those features that are particular to a character, thereby providing a robust method for the classification of features that is invariant to the different shapes or sizes of the characters. We then create a database, where we store the chain codes for these lower level features. Now, when an input character is to be identified, we first find out what basic features does it have. This can be done using simple techniques of calculating row and/or column densities, pixel connectivity. Depending on the path in the hierarchy that the character follows, a lower level specific feature is identified. This is done using chain code techniques, which follows the shape of the character. In this method, we propose that the hierarchy should be followed in strict order. If a match with the first group is not found, then algorithm should proceed to the next group in the same hierarchical level. However, if a match is found in one group, the algorithm should proceed down to the next hierarchical level within that group. This paper is outlined as follow. Section 2 presents the hierarchical classification of features. Section 3 describes the database which contains chain code of the features with which the input character is to be matched. Section 4 presents the stepwise algorithm for our method. Section 5 describes the procedure in details followed by an illustration. Section 6 shows the different types of test inputs followed by the results and discussions. Section 7 gives a conclusion and future research scope of our work.

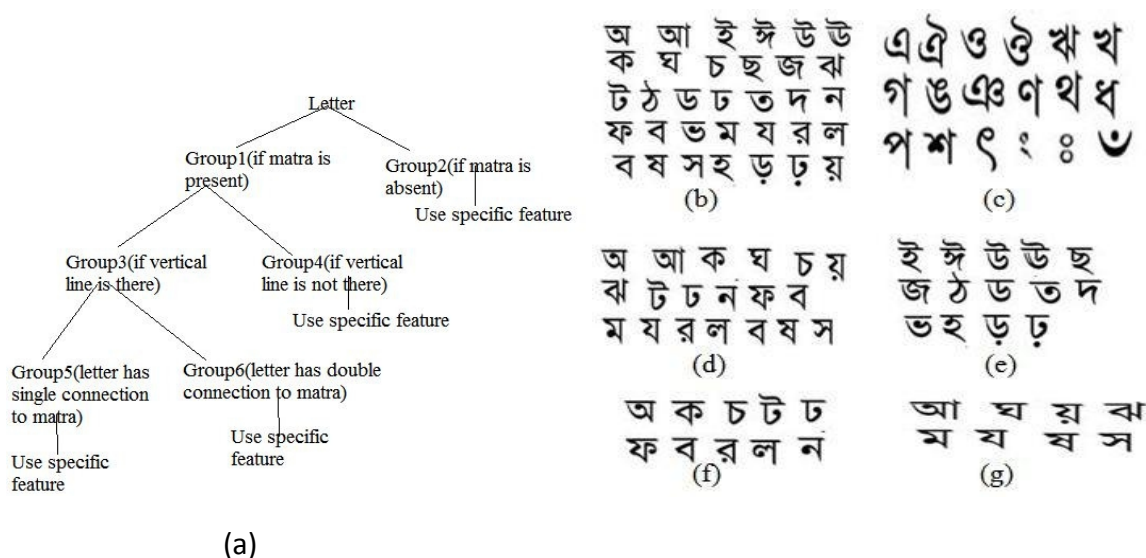


Figure 1: (a): Hierarchy, (b)-(g): Groups 2-6

Figure 1 (a) represents the hierarchical classification scheme. Figure 1(b) depicts the presence of “matra” labelled as Group 1, while Figure 1 (c) represents Group2 - absence of “matra”. Figure 1 (d) denotes Group3 : presence of vertical line, while Figure 1(e) is Group 4- absence of vertical line. Figure 1(f) is Group 5 which has single connectivity to “matra” and Figure 1(g) and Group 6 shown in Figure 2 denotes double connectivity to “matra”.

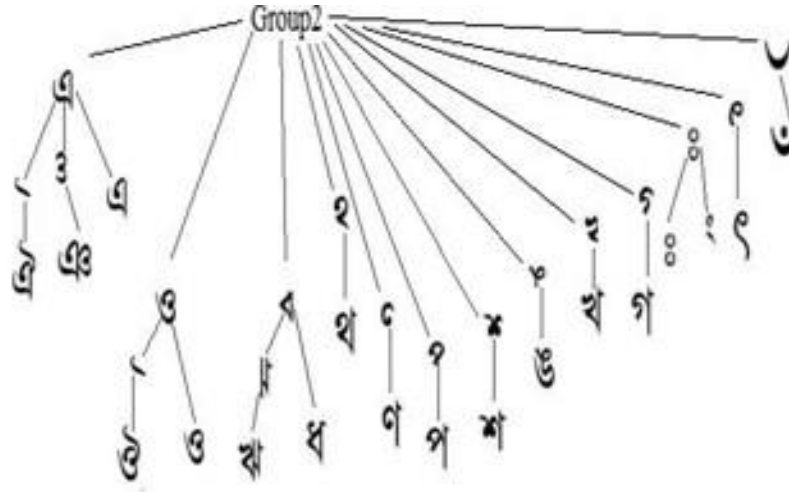


Figure.2: Group 2

After further subdivision, classification based on specific features is shown in Figures 2-5 with Figure 2 labeled as Group 2, Figure 3 is Group 4 , Figure 4 is Group 5 and Figure 5 is Group 6. All these groups have been further subdivided based on specific features as depicted in figures.

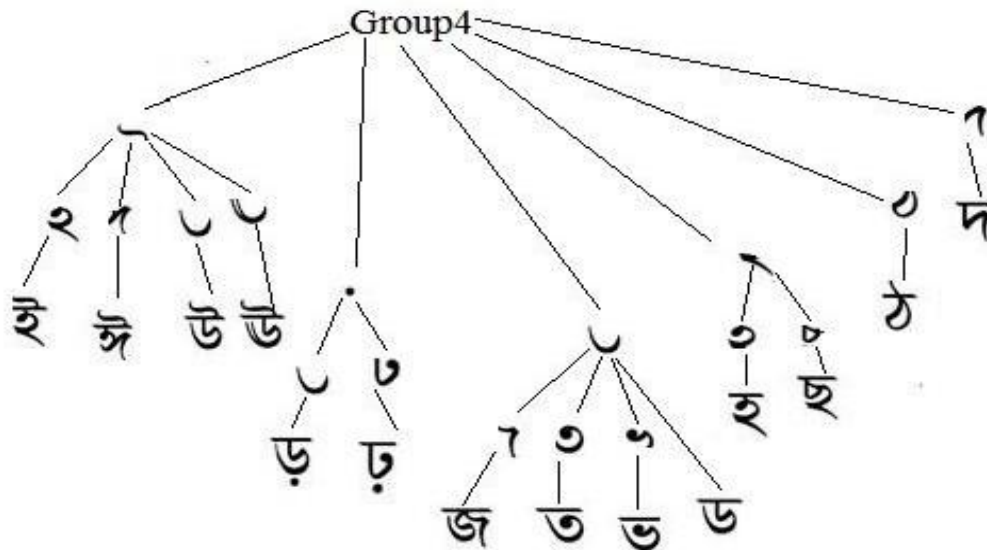


Figure 3: Group 4

2 FEATURE EXTRACTION

Feature extraction is the crucial first step of character identification in our proposed algorithm. Classification starts with the detection of the “matra” (the horizontal headline over some of the characters) in the character. Based on this detection,,group1 and group2, respectively, are defined.

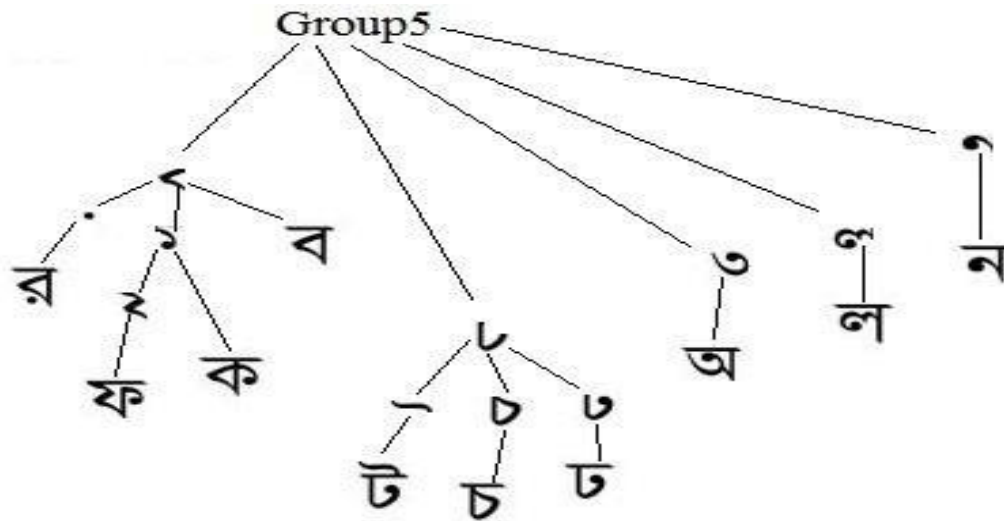


Figure 4: Group 5

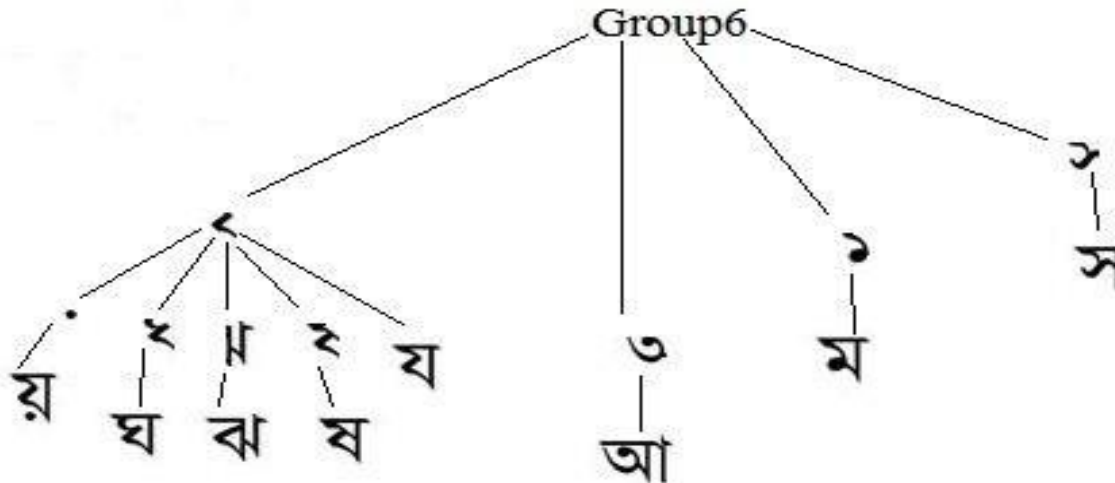


Figure 5: Group 6

The characters in group 1 can be further subdivided into group3 and group4 based on the presence or absence of a vertical line either in the beginning or at the end of the character. Further, the characters in group3 can be subdivided based on the number of places where the “matra” is connected to the character below it. Then, group 5 is defined representing characters having single connectivity to “matra” and group 6 is defined representing characters having

double connectivity to the “matra”. Thus, the basic features which divide the character set into similarly sized groups at each level of the hierarchy have been identified. Now, the characters in the groups which are at the leaf level will need to be identified based on specific features of the particular character. Thus, for each character, features exclusively identifying the character within that group have been defined. The features used to classify each of group 1 to group 6 are labelled as the higher level basic features and the features used in the rest of the hierarchy are labelled as the lower level specific features. The full classification is shown in Figure 6.

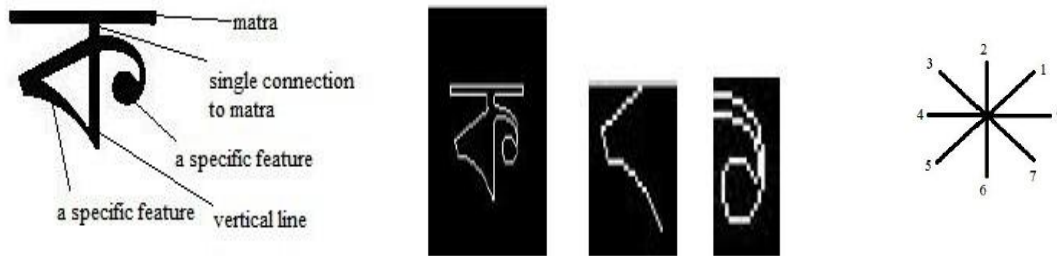


Figure 2: Steps of the method

3 DATABASE

The lower level specific features discussed in section 2 are such features that can directly identify which character it represents within a particular subgroup. In order to train the computer to identify these features in a given input image, the chain code representation for each such feature is determined. The chain code is obtained from the contour representation of the feature and is stored in the database. This representation is to be used later to find a match with the character’s chain code representation.

4 ALGORITHMS

The algorithm is presented as follows:





1. Obtain the input character in RGB form and scale the image to a predefined size.
2. Use Otsu’s method to find out the global threshold for the image.
3. Using this threshold, convert the image to logical form.
4. If the background pixels are white, that is represented by logical ‘1’, then complement the image so that the background pixels are represented by logical ‘0’, else go to step 5.
5. Check to see if the character has a “matra” or not. If yes, then put it in group 1 and proceed to step 6 else put it in group 2 and proceed to step 8.

6. Check to see if the character has a vertical line in the beginning or end of the character or not. If yes, then put it in group 3 and proceed to step 7 else put it in group 4 and proceed to step 8.
7. Check to see if the character is connected to the “matra” at one point or at two points. In the former case, put it in group 5 and proceed to step 8 and in the latter case, put it in group 6 and proceed to step 8.
8. Now, the character could be in either of group 2, group 4, group 5 or group 6. Find out the chain code for the contour of the character.
9. For each group, check in order as shown in the feature classification hierarchy, if the chain code of the features for that group which is stored in the database are found in the chain code for the character contour obtained in step 8.
10. If a match is found then proceed downwards to the group in the next hierarchical level until the character is identified and go to step 11 else proceed to the next group in the same hierarchical level to find out if the character can belong to that group and go to step 9.
11. Algorithm ends.

5 METHODOLOGY

When the input image is obtained in the RGB format, scaling operations on the image are first performed so that the image is of the standard size which has been used in this method. This is followed by converting it to logical form by using Otsu’s global threshold method [7]. If necessary, the complement of the image is found so that the background pixels are represented by ‘0’ and the foreground pixels are represented by ‘1’. Now, the character is identified. Following the hierarchical order, a check is made to see if the character has a “matra” or not. This can be checked by the fact that the identified rows in the image corresponding to the “matra” will have a relative density greater than or equal to 70%. Thus the character can fall in either group1 or group 2 depending on whether the “matra” is present or not respectively.

Now, for characters in group 1, a check can be made to see if there is a vertical line in the beginning or end of the character. This can similarly be checked, because the columns representing such a line would have a relative density greater than or equal to 70%. Thus the character can fall in group 3 or group 4 depending on the presence or absence of the vertical line respectively. The characters of group 3 can be further checked to see the connectivity to the “matra”. The character below the “matra” is joined to it either at one point or two points. The width of the connection is also very small, less than 5% of the total number of pixels in the row. Thus the character can fall in group 5 or group 6 depending on whether the character has a single connection to the “matra” or double












The second feature at the same level is . The chain code for this feature is 0000070077077667 6666 6656555 44444323222 2110000777 122232334 34434 444. Now this pattern can be found in the chain code for the character and a match is found. Then proceed downwards into the subgroup. The first feature encountered is . The chain code representation for this feature is 000007776555570707634344 34231 101133443. However, such a pattern in the chain code for the character  is not found and since there are no other options left, therefore this character has to be the other one in that subgroup and so this method has correctly identified the character as .

6 RESULT AND DISCUSSIONS

A large number of samples for each character have been collected [9] and tested using the proposed method. The results for the experiment are given in Table 1.

It is seen from the results, that the chain code matching algorithm gives a high accuracy of matched characters. The chain code follows the direction that the characters take, encoding the shape of the desired feature. Using a variety of images, where the characters may be represented in various fonts and sizes, this approach gives satisfactory results because the chain code will follow the direction of the feature, which will always be same for all cases. The negligible differences that occur have been studied carefully and accounted for. Although Chain code matching technique gives very good results in most of the cases, it is cumbersome to do so for each and every feature and write algorithms for each such feature.

Table 1. Results of the method.

LETTER TAKEN	NUMBER OF SAMPLES TAKEN	PERCENTAGE OF MATCH
	50	98
	50	100
	50	94
	50	92
	50	94
	50	94
	50	94
	50	100
	50	94
	50	98
	50	90

7 CONCLUSION AND FUTURE WORK

Based on the results obtained, it can be concluded that the classification accuracy has been over 90 %, with the best three letters having an accuracy of 100% and 98%. In some cases a character may have been misclassified because the font used did not represent the character in its proper format. It is difficult to account for all the different fonts available for writing Bengali characters. However, the feature classification presented in this paper is based on crucial features that have almost similar representations in every font system. Therefore, this classification is quite robust. The algorithms used to match with the patterns are quite flexible. Although many machine learning based approaches to Bengali character recognition are being attempted, this simple method has its strength in identifying the characters based on certain very crucial characteristics and is thus a very universal approach, which can be extended to Bengali handwritten character identification in the future.

REFERENCES

- [1]. Mohammed Jasim Uddin, Mohammed Towhidul Islam and Md. Abdus Sattar, *Recognition of Printed Bangla Characters Using Graph Theory*, National Conference on Computer and Information System-NCCIS, Dec 9-10, 1997, Dhaka, Bangladesh
- [2]. Chaudhuri, B. B., Pal, U.: A Complete Printed Bangla OCR System. *Pattern Recognition*, Vol. 31. (1998) 531-549
- [3]. Ujjwal Bhattacharya, Malayappan Shridhar, and Swapan K.Parui. On recognition of handwritten bangla characters. In *ICVGIP*, pages 817- 828, 2006.
- [4]. J.U. Mahmud, M.F. Raihan and C.M. Rahman, "A Complete OCR System for continuous Bengali Character", *TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region*, 15-17 Oct. 2003
- [5]. Dewi Nasien, Habibollah Haron, Siti Sophiyati Yuhaniz, "The Heuristic Extraction Algorithms for Freeman Chain Code of Handwritten Character", *International Journal of Experimental Algorithms-IJEA*, Vol. 1, Issue 1, pages 1-20.
- [6]. Trier, O. D., Jain, A. K. and Taxt, T.: Feature Extraction Methods for Character Recognition - A Survey. *Pattern Recognition*, Vol. 29 (1996) 641 - 662
- [7]. Otsu, N.: A Threshold Selection Method from Grey-Level Histograms. *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 9 (1979) 377-393
- [8]. Freeman, H.: Computer processing of Line-drawing Images *ACM Computing Surveys*, Vol. 6 (1974) 57-97
- [9]. Sikdar A., Roy P., Mukherjee S., Das M. and Banerjee S., A Feature Based Chain Code Method for Identifying Printed Bengali Characters, (2012) *Proceedings, SIPM 2012*, 89-96.