A FEATURE BASED CHAIN CODE METHOD FOR IDENTIFYING PRINTED BENGALI CHARACTERS

Ankita Sikdar[1], Payal Roy[1], Somdeep Mukherjee[1], Moumita Das[1] and Sreeparna Banerjee[2]

[1]Department of Computer Science and Engineering, West Bengal University of Technology, Kolkata, West Bengal, India
ankita.sikdar@gmail.com
payalroys@gmail.com
somdeep.mukherjee@gmail.com
moumitadas8484@gmail.com
[2]Department of Natural Sciences, West Bengal University of Technology, Kolkata, West Bengal, India
sreeparnab@hotmail.com

ABSTRACT

This paper gives complete guidelines for authors submitting papers. This paper aims to explore a new way for recognizing printed Bengali characters. Keeping in mind, the possible shapes and orientations of the Bengali characters, we have developed a method to classify each of the 50 Bengali characters. An exhaustive study of the features of Bengali characters has been carried out which is presented in 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. While the higher level features can be identified based on pixel density and arrangement, the lower level features have been identified using chain code technique. The computer has been programmed to progress successively through each group in the hierarchy until it finds a match with the input character or rejects it.

KEYWORDS

Bengali Character Feature Identifications, Chain Code Technique & Feature Extraction

1. INTRODUCTION

Bengali is one of the official languages of India. It is widely used in the states of West Bengal, Tripura and Assam. Also, Bengali is the official language of neighboring country, Bangladesh. The Bengali alphabet consists of 50 characters, out which there are 11 vowels and 39 consonants. Therefore, identifying these printed characters is important in applications where we have to digitize the printed text in order to create an online version of the text or in image processing applications where 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 lot of research work has been carried out to identify printed characters and a discussion has been provided in [2]. Chain code has been used in our method. Some of the other ways in which chain code techniques have been used are presented in [3,4,5].These are based on feature

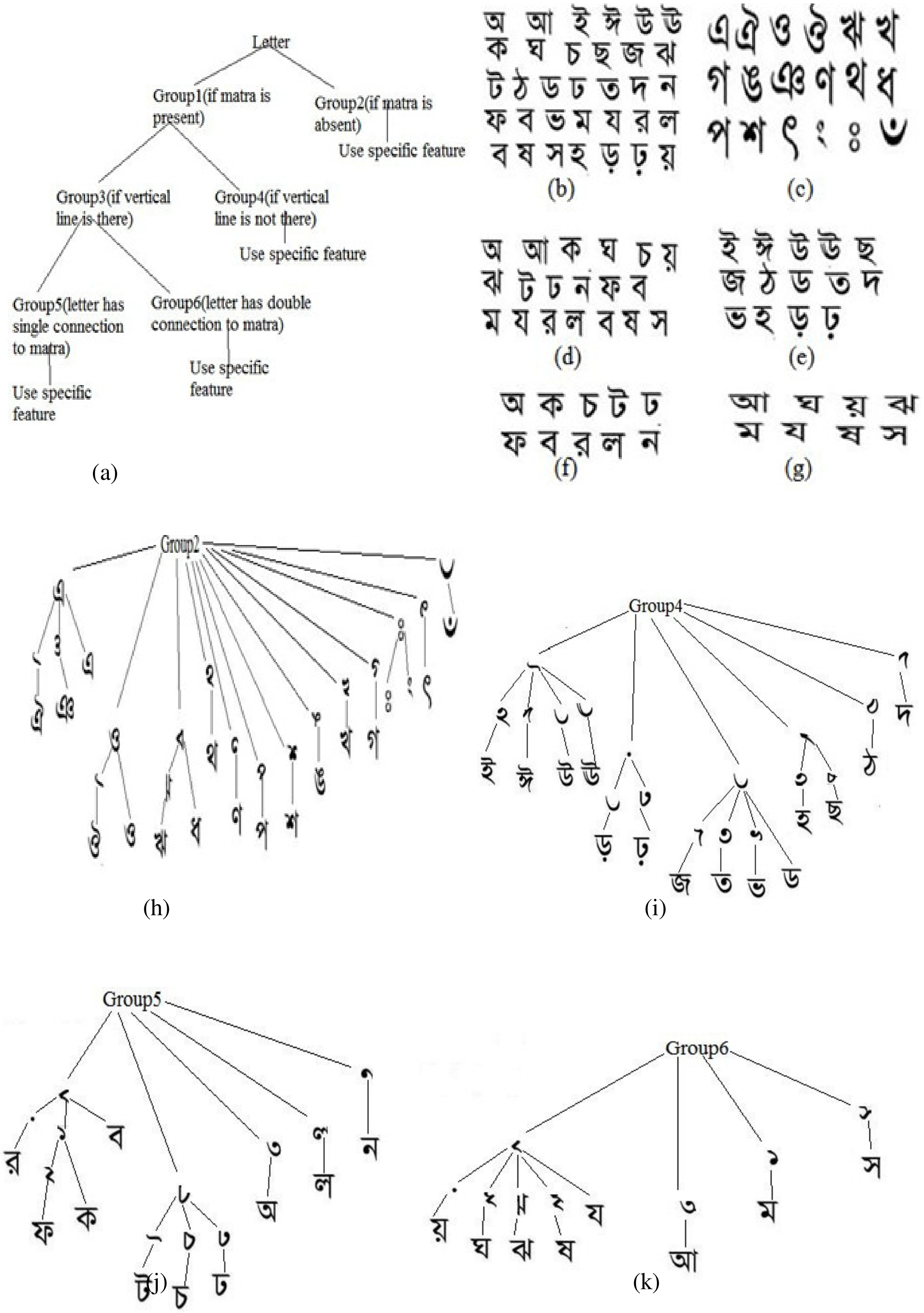
Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 89–96, 2012.

DOI : 10.5121/csit.2012.2310

extraction followed by various methods to use these features to identify the characters. A comprehensive study for feature extraction has been presented in [6]. In this paper, we present a different method. In our aim to identify these 50 characters, we have first designed a feature set that introduces us to the features of the characters in a hierarchical manner, 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 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 follow 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 along with 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 on our work.

2. FEATURE EXTRACTION

Feature extraction is the heart of identifying characters. It is quite challenging to design a feature set that will be able to cater to an exhaustive sample set of Bengali characters of varying shapes, sizes and orientations. Our first classification is based on the presence or absence of the “matra” (the horizontal headline over some of the characters) in the character. These form group1 and group2 respectively. Now, 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 end of the character respectively. Further, the characters in group3 can be subdivided based on the number of places where the “matra” is connected to the character below it. So, we have group5 which represents characters having single connectivity to “matra” and group6 which represents characters having double connectivity to the “matra”. So, now we have identified the basic features which divide the character set into similarly sized groups at each level of the hierarchy. 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 now, we have found out such feature exclusively identifying the character within that group. We call the features used to classify each of group1 to group6 the higher level basic features and the features used in the rest of the hierarchy as the lower level specific features. The full classification is shown in Figure1.



(a) (b) (c) (d) (e) (f) (g)

(h) (i)

(j) (k)

Figure 1. Feature classification

(a) Hierarchy
(b) Group1 (presence of "matra")
(c) Group2 (absence of "matra")
(d) Group3 (presence of vertical line)
(e) Group4 (absence of vertical line)
(f) Group5 (single connectivity to "matra")
(g) Group6 (double connectivity to "matra")
(h) Group2 subdivided based on specific features
(i) Group4 subdivided based on specific features
(j) Group5 subdivided based on specific features
(k) Group6 subdivided based on specific features

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, we find out, for each such feature, its chain code representation. 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. ALGORITHM

The algorithm for our method 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 group1 and proceed to step 6 else put it in group2 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 group3 and proceed to step 7 else put it in group4 and proceed to step8.

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 group5 and proceed to step8 and in the latter case, put it in group6 and proceed to step8.

8. Now, the character could be in either of group2, group4, group5 or group6. Find out the chain code for the contour of the character.

9. For each group, check in order as shown in 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. PROCEDURE

When the input image is obtained in the RGB format, we first perform scaling operations on the image so that the image is of the standard size which we have used in our method. This is followed by converting it to logical form by using Otsu's global threshold method [7]. If necessary, we also find complement of the image so that the background pixels are represented by '0' and the foreground pixels are represented by '1'. Now, we proceed to identify the character. Following the hierarchical order, we first check if the character has a "matra" or not. This can be checked by the fact that the 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 group2 depending on whether the "matra" is present or not respectively. Now, for characters in group1, we can check 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 group3 or group4 depending on the presence or absence of the vertical line respectively. The characters of group3 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 less, less than 5% of the total number of pixels in the row. Thus the character can fall in group5 or group6 depending on whether the character has a single connection to the "matra" or double connections respectively. Now, group2, group4, group5 and group6 represent character sets based on the commonly observed features of Bengali characters.

Up to this point, we have identified in which broad group the character can fall. Next, we proceed to identify what character it is. For this purpose, we have selected features which are typical of a particular character and in most cases this single feature either identifies the character directly or identifies a group of two to four characters at most. For every such feature, we have found out the chain code [8] and stored these in a database. Now, for the character, we will find out its entire chain code and then see if in this chain code, there exists a pattern that corresponds to the chain code for the feature of that group. We perform this matching by trying to identify in the chain code for the character, certain groups of strings that are specific to the feature that we are trying to match with. If a match is found, then we proceed downwards in the group until an exact match with a character has been found. If a match is not found, then we conclude that that character does not belong to that group and we proceed to the next group in the same hierarchical level.

Figure 2 illustrates the proposed method with the help of an example. An input character ক is given and the various steps that it goes through are shown. After the pre-processing steps involved, we proceed to identify the given character. At first, it is determined that ক has a "matra" and so it falls in Group1. This is followed by checking for the presence of a vertical line. Since it is present, ক falls in Group3. Also, ক is connected to the "matra" at a single point and so according to our proposed classification, the character ক falls in Group5. Now, we have to identify particular features in ক. We now obtain the contour representation for ক which is shown in Figure 2(b). Using this contour representation, we obtain the chain code. The chain code for this contour of ক is

00000000000000000000000000000000000000000000000000000666444444444444444444444456
67007000070077077667666666565554444432322221100007771222323343443444444556666666
66666666666666666666423333333334343344343444444323232010010101010010101010010112
2234444444444444444444444444444222

The first specific feature given for Group5 is <. The chain code representation for this feature is 3333333434334434344444432323201001010101001010. It has to be kept in mind that the chain code for the features has to be obtained in the same direction in which the contour of the character to be matched is traced for error-free results. So here the chain code of < is traced from the bottom. Since this pattern for < is found in the pattern for ক , therefore we proceed downwards into the next hierarchy. The first distinguishing feature now for this sub group is • . The chain code for • is 11221110100070767666666565454444443332232.This pattern is not found in the chain code representation for ক and so we move on to the next group in the same level of the hierarchy. The second feature at the same level is ɔ . The chain code for this feature is 0000070077077667 6666 6656555 44444323222 2110000777 122232334 34434 444.Now

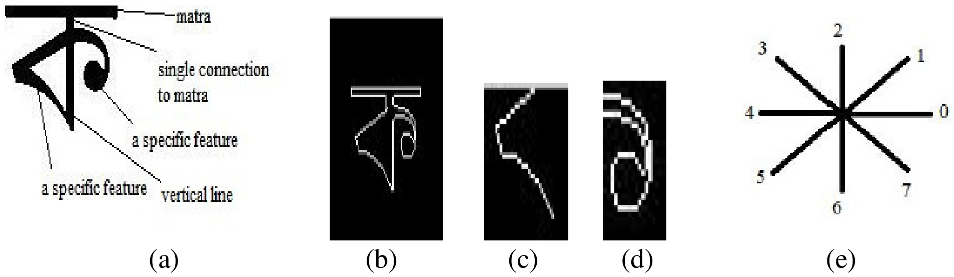


Figure 2. Steps of the method

(a) A character
(b) Contour representation
(c) Specific feature for that character
(d) Specific feature for that character
(e) 8- directional chain code

this pattern can be found in the chain code for the character and so we find a match. Thus we proceed downwards into the subgroup. The first feature we encounter now is ২ . The chain code representation for this feature is 000007776555570707634344 34231 101133443. We do not however find such a pattern in the chain code for the character ক and since there are no other options left, therefore this character has to be the other one in that subgroup and so our method has correctly identified the character as ক.

6. Results and Discussions

We collected a large number of samples for each character and tested these samples with our proposed method. The results for the experiment are given in Table1.

Table 1. Results of the method.

LETTER TAKEN	NUMBER OF SAMPLES TAKEN	PERCENTAGE OF MATCH
ক	50	98
র	50	100
ঘ	50	94
ই	50	92
উ	50	94
ত	50	94
অ	50	94

ঐ	50	100
ণ	50	94
ও	50	98
ড	50	90

As we can see from the results, the chain code matching algorithm gives us a high percentage of matched characters. The chain code follows the direction that the characters take, encoding the shape of the desired feature. When we get 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, yet it is very laborious to carry on an exhaustive study of each and every feature and write algorithms for that.

7. CONCLUSION AND FUTURE SCOPE

Using our proposed method, the printed Bengali characters have been identified successfully in a large number of cases. In some cases where 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 have to be quite flexible. In the future, we hope to extend our work of feature classification and chain code matching to handwritten Bengali characters where it would also prove to be very beneficial.

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 Sophiayati 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

Ph

Authors

Ankita Sikdar has done her schooling form Mahadevi Birla Girls' Higher Secondary School. She is at present a fourth year student of West Bengal University of Technology pursuing B.Tech in Computer Science and Engineering. She is going to pursue Phd in The Ohio State University, research area : Artificial intelligence

Payal Roy has done her schooling from Carmel School and Hem Sheela Model School. She is at present a fourth year student of West Bengal University of Technology,Kolkata pursuing B.Tech in Computer Science and Engineering. She has bagged a few job offers and will be soon joining the industry.

Somdeep Mukherjee has completed his schooling from St. Xavier's Collegiate School, Kolkata. He is at present a fourth year student pursuing B.Tech in Computer Science and Engineering from West Bengal University of Technology, Kolkata. He has got job offers and is about to join the industry.

Moumita Das is a fourth year student of West Bengal University of Technology pursuing B.Tech in Computer Science and Engineering. She is about to join the industry.

Dr Sreeparna Banerjee obtained her B. Sc., M.Sc., and Ph.D degrees all in Physics. She has taught in universities in India and abroad. Her current research interests include Physics of space plasmas: Molecular Dynamics and Monte Carlo simulations,charge transfer, nonlinear dynamics; Neural networks, Pattern Recognition and Soft Computing applications in Astrophysics, Meteorology and Medical Imaging; Data Mining.