Analysis of Different Techniques for Segmentation of Connected Handwritten Indian Scripts Words

Dharam Veer Sharma[1], Supreet Kaur[2]

Department of Computer Science, Punjabi University, Patiala

dveer72@hotmail.com[1], ghumansupreet@gmail.com[2]

Abstract: *In OCR processing text is extracted from an image. Here many common imperfections occur, like the characters that are connected together are returned as a single sub-image containing both characters. So this is the major challenge in the recognition process and we need to segment the connected words to get correct results. In segmentation process the first task is to find out the segmentation point. Once this is done the next step is to fragment the image from that point and get two separate sub-images for each character. There are many algorithms proposed for this process. One class of approaches uses the straight segmentation technique and segments the words based on the vertical and horizontal histogram profiles. Another approach uses contour features of the component for segmentation. Some researchers segments the characters by recognizing the profile features of touching character and the segmenting them. This technique is commonly known as recognition based segmentation or cut classification.*

1. Introduction

In Optical character recognition (OCR) processing, the scanned-in image is analyzed for light and dark areas so as to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code or any other digital code. OCR technology has made a huge

impact on the way information is stored, shared and edited. Before the development of OCR technique, if someone wanted to turn a book into a word processing file, each page would have to be typed word for word.

There are various OCR (Optical Character Recognition) tools available for extracting texts from images. This process involves many steps for recognizing printed or handwritten text. First of all image needs to be scanned. It is followed by analysis of the scanned image and then translation of the character image into character codes, such as ASCII, commonly used in data processing. In this process, a symbolic meaning is associated with objects drawn on an image. The ultimate goal of OCR is to imitate the human ability to read at a much faster rate by associating symbolic identities with images of characters.

Segmentation is a crucial step of OCR systems as it extracts meaningful regions for analysis. In OCR systems when text is extracted from an image, several common imperfections occur. The most common being characters that are connected together are returned as a single sub-image containing both characters. So this is the major challenge in the recognition process. To overcome this imperfection we need to segment the connected words. In segmentation process the first task is to find out the segmentation point. Once this is done the next step is to fragment the image from that point and get two separate sub-images for each character. However the algorithm for segmentation should produce correct results i.e. the point of segmentation should be chosen such that the results are not erroneous.

Segmentation is a very essential component in any practical handwritten recognition system. This is because handwriting is unconstrained and depends on writers. It is commonly noticed that whenever we write in any script, in our day-to-day life, we tend to connect the characters. There are many algorithms proposed for this process. One class of approaches uses the straight segmentation technique and segments the words based on the vertical and horizontal histogram profiles. Another approach uses contour features of the component for segmentation. Some researchers use profile features for touching character segmentation this technique is commonly

known as recognition based segmentation or cut classification. Some of these methods have been discussed in the following section.

The main focus while studying these techniques is the Indian scripts like Devnagri, Gurmukhi and Bangla. This is because these scripts have common structural features. Some of these features are.

i. The writing style for these scripts is from left to right.
ii. Most of the characters have horizontal line at the upper part. This line called headline, and it connects the characters of the word. Normally, vertical inter character gaps are not present in the letters of word.
iii. A word in these Indian scripts can be partitioned into three horizontal zones
    - o The upper zone denotes the region above the headline containing vowels, so identification of headline solves the purpose of finding vowels above headline.
    - o The middle zone denotes the area below the headline where consonants and some subparts of vowels are present.
    - o The lower zone denotes the area below middle zone where some vowels and certain half characters lie in the foot of full character. Usually this area has the minimum density.
iv. A vowel in a word is always present with a consonant i.e. it cannot exist on its own.
v. The character in lower zone sometimes touches the character in middle zone
vi. There is no concept of upper/lower case is absent.

2. Analysis of Segmentation Techniques

Straight segmentation method: Each word is segmented into several characters and then a character recognition technique is applied to each segment. It is simple but it depends on the accuracy of detection of segmentation points. It is known as Classical or dissection approach. For most Indian scripts like Gurmukhi, Devnagri & Bangla this technique shows good results. First of all horizontal histogram is computed to detect the presence or absence of headline. It also helps to identify the position of the

headline of the word. Having found the location of headline, divide the word in two horizontal zones, one just above the headline and the other below the headline. This division will make gaps in the characters of the word which are segmented by creating vertical projection profiles. But if the headline is not identified because of its absence then vertical histogram is calculated to identify gaps for segregation. To segment the lower zone, the presence of any gap in the lower part (threshold aspect ratio is used for this calculation) of the horizontal histogram of zone below headline is found out. If no gaps are found up to a threshold then either the lower zone characters are not present or they may be overlapped or connected with the middle zone characters. Once all the zones are detected, the middle zone characters are then segmented at the position of gaps in vertical histogram (computed by considering the area just below the headline as the top starting point). This step gives us the number of columns formed in the word, which helps in associating characters in the upper and lower zone with characters in the middle zone.

In next phase under segmentation is handled. For middle zone characters a threshold value of aspect ratio is used to identify whether the characters have been segmented properly or they are under segmented. If the aspect ratio of any segment is higher than the threshold value of aspect ratio, then there may be two or more connected characters present in that segment. For connected character again the procedure of horizontal histogram is applied, ignoring the headline. The hill and valley is detected in the histogram from minimum value of histogram. This is then used to segment the connected characters [1, 2].

Recognition-based segmentation method: In recognition based segmentation technique once we have an image where characters are connected or merged; it is ready for the feature extraction stage. The fragments of this image are created and then these fragments are used to extract the structural features of each character fragment. Once we have features extracted by this method, these fragments are the recognized by the classifier. Based on the result of classification these connected words are segmented. This method is fully dependent on the performance of the recognizer.

This method is also known as Cut classification method. This is because it is based on a classifier deciding whether it represents a cut hypothesis or not, for each column of the character image. This strategy intends to recognize words as entire units rather than attempting to extract individual characters. It is known as holistic approach.

The most commonly used feature for detecting connected character segments is the width and the aspect ratio of the character. Each segment is examined by comparing its width with the estimated character width (this estimation is completely dependent on the script and is estimated by experimentation) or by measuring the aspect ratio of the segment. It is well understood that a single character segment should have a width of less than the estimated character width. This measurement works well most of the time but sometimes fails in special cases.

Once the candidate for segmentation is selected in above step, then the fragments of that segment are created. For each fragment the structural features are extracted and it is then recognized using a classifier. The point of segmentation is detected based on these results.

For selecting the point of segmentation different algorithms have been proposed. One approach constructs the binary tree based on the structural features of the script. For recording the structural features it uses the Freeman code. Each node in the tree describes the corresponding part of the word. Segmentation involves splitting the binary tree into number of sub trees, each representing a separate character [3]. Another approach assigns an input character to one of many pre-specified classes which are based on the extracted features and their analysis. The above processes are repeated until a character is recognized. If a character is recognized after the combination of the first n fragments, then the feedback loop will start again at the (n+1) th fragment. The above feedback loop occurs twice for each word. The first is with the fragment combination directed from right to left of the word. If not all characters in that word are recognized, the second feedback loop proceeds. This time, the fragment combination is directed from left to right of the word. The results from these two feedback loops are combined to form the final recognition results [4].

Contour Tracing: Contour Tracing is also known as *border following* or *boundary following*. Contour tracing is a technique that is applied to digital images in order to extract their boundary. Contour tracing is one of many preprocessing techniques performed on digital images in order to extract information about their general shape. Once the contour of a given pattern is extracted, its different characteristics are examined and used as features which are later on used in pattern classification. Therefore, correct extraction of the contour will produce more accurate features which increase the chances of correctly classifying a given pattern. In this project, we are using Contour tracing to segment overlapping characters or symbols.

Two most common Contour Tracing Algorithms are:

1. Square Tracing Algorithm: For a given digital pattern, i.e. a group of black pixels, on a background of white pixels known as a grid, locate a black pixel and declare it as the "start" pixel. The location of a "start" pixel can be done in a number of ways. In approach for Indian scripts would, the process starts at the top left corner of the grid. Then the pixels are scanned diagonally going right-downwards until a black pixel is encountered. It is declared as the "start" pixel. Then for tracing turn right on black pixels and turn left on white pixels. The tracing process stops when start pixel is encountered again.
2. Moore-Neighbor Tracing Algorithm: Locate a black pixel and declare it as the "start" pixel. Every time a black pixel is hit, backtrack i.e. go back to the white pixel you were previously standing on, then, go around current pixel in a clockwise direction, visiting each pixel in its Moore neighborhood, until a black pixel is hit. The algorithm terminates when the start pixel is visited for a second time. The black pixels that are walked over will be the contour of the pattern.

The contour tracing step is very important because during this step only connected characters are segmented. One technique uses recursive contour following to compute the segmentation point. The image is scanned from left to right in the zone bounded up by headline and below by the word boundary. The first object pixel hit is recorded. Starting with this pixel, a recursive contour following is done in the said zone. To find the merging of character extents it is seen that if the convex hull of any character extent is

contained in another convex hull, the smaller one is merged with the larger one and if the y-value of the lower bound of any character extent is less than threshold, this character extent is merged with the one to its immediate right and the character extent values are updated. This way the point of segmentation is detected and they are separated [5].

Water Reservoir Principle: The water reservoir principle states that if the water is poured from top/bottom then where water will be collected is called as the top/bottom reservoir. In first step lines and words are separated by using piece-wise projection method. Then the process to segment characters from the words begins. From water reservoir principle we can say that where two Bangla characters touch, two consecutive characters create a large bottom reservoir or the number of reservoirs in a connected component will be greater than that of an isolated component or vertical overlapping of a reservoir and a loop or vertical overlapping of two reservoirs does not exist in handwritten Bangla isolated characters whereas in connected character such overlapping occurs frequently For each reservoir a base area is computed. To segment the characters from the word first connected and isolated characters are found out. Connected characters are recognized by their looking at their bottom reservoir which is exceptionally large or number of reservoirs is large. Once the connected characters are found out then they are separated from the point of touch based on touching position, reservoir base area and structural features. To segment a connected component into two components, cutting is done vertically at the best feature point in the direction opposite to the reservoir. Two segmented parts are then passed to the isolated and connected character detection module to check whether any of these segmented parts is connected. If any part is detected as connected it is then sent to segmentation module for further segmentation. This procedure is repeated until both the segmented parts of a component are detected as isolated by the isolated and connected character detection module [6].

3. Conclusion

Though the different segmentation techniques discussed above have shown remarkably good results, but each has its own pros and cons. The

straight segmentation technique is simple and easy to implement. It works efficiently for segmenting handwritten words of Indian scripts except for connecting and touching characters. Moreover it is dependent on many heuristics e.g. the presence of valley and hill is considered as point of segmentation. Recognition based segmentation technique shows highly good results but it is highly dependent upon the feature extraction and classification stage. If a character in a word is incorrectly recognized, the rest of the characters in the word will not be recognized properly as well. It seems, by offering a better classification method, this problem could be solved. The contour tracing method is comparatively more efficient but it is highly dependent upon the structural features of the script. Moreover it could lead to over-segmentation of the words. The advantage of water reservoir principle is that it is size independent and there is no need any normalization of the component. Here the segmentation of multi-touching points between two characters did not show significant results. Each of the above algorithms has been successfully implemented on Indian scripts by various researchers.

References

[1] Dharam Veer Sharma and Gurpreet Singh Lehal, “An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script”, in proceedings of the 18th International Conference on Pattern Recognition, ICPR-2006, Hongkong, Vol. 2, pp. 1022-1025 , August, 2006.

[2] Ashraf Elnagar, Reda Alhajj “Segmentation of connected handwritten numeral strings”, in proceedings of JOURNAL = “Pattern Recognition”, Vol. 36, Issue 3, pp. 625-634, March 2003.

[3] Adnan Amin, Humoud B. Al-Sadoun “A new segmentation technique of Arabic text”, in proceedings of the 18th International Conference on Pattern Recognition, Vol. 2, pp. 441-445, 1992.

[4] Mohsen Zand, Ahmadreza Naghsh Nilchi, and S. Amirhassan Monadjemi, “Recognition-based Segmentation in Persian Character

Recognition", in proceedings of International Journal of Computer and Information Engineering 2:5 2008.

[5] A. Bishnu, B.B.Chaudhri "Segmentation of Bangla Handwritten Text into characters by recursive contour following", in proceedings of 5th International Conference on Document Analysis and Recognition (ICDAR '99), pp. 402-405, 1999.

[6] U. Pal, Sagarika Datta "Segmentation of Bangla Unconstrained Handwritten Text", in proceeding of 7th International Conference on Document Analysis and Recognition (ICDAR '03), Vol. 2, pp. 1128-1132, Aug-2003.