Machine Printed Punjabi Character Recognition Using Morphological Operators on Binary Images

Usha Rani
Computer Science Deptt.
Punjabi University Guru Kashi College,Damdama Sahib

Er. Balwinder Singh
Computer Science Deptt.
Govt. Multipurpose Sec. School, Patiala

Er. Ravinder Singh
Computer Science Deptt.
Govt. Multipurpose Sec. School, Patiala

ABSTRACT

In this paper we present a character recognition system by using morphological operators on binary images. As a consequence, we will deal with the Punjabi language characters. This recognition system is merely feature-based, with no need of a learning phase or any kind of memory. Main advantage of this system is its accuracy to recognize Punjabi characters. Input to the system is the scanned images from newspaper, magazines and old books.

Keywords: Morphological Operators, Punjabi character recognition, machine printed, binary images .

1. Introduction

Optical Character Recognition is a technology used to copy and machine printed material into editable word processing file formats. This is the technology long used by libraries and government agencies to make lengthy documents quickly available electronically.
Optical character recognition is a challenging problem, the solution of which the researchers have been pursuing since more than 100 years. Several algorithms have been proposed to improve recognition capabilities [1,2]. Methods used to recognize characters inside a bitmapped image fall mainly into two categories: pattern matching, used in cheaper systems, and feature analysis, used in more sophisticated systems.
Pattern matching methods have the bitmaps stored for every character of each of different font and type sizes. By comparing a database of stored bitmaps with the bitmap of scanned character the program tries to recognise the letters. The pattern having best correlation is considered to be the scanned letter. If none of the pattern has a sufficient high degree of correlation with the scanned character, the character is considered to unclassifiable. This family of methods is usually not computationally heavy as well as quite robust to noise. But this system has the severe drawback: that is the lack of generality as it is only useful for the fonts and sizes stored. Complex multifont documents are beyond its scope. On the other side, feature extraction method attempts to recognise characters by identifying their universal features and make OCR type –face independent. In this method optical scanner looks for certain features in letters such as intersections in lines, diagonal lines, shapes in the characters that are closed, shapes that are open etc. Then these read features of scanned letter are compared to list of features that are available in the software‘s programming code. This method is more versatile because it works with many types of fonts and characters.
Our algorithm is feature analysis based. As a consequence, it shows great generalization capabilities. Moreover, all needed features can be extracted by using Morphological Operators.
Morphology is the branch of biology that deals with the form and structure of animals and plants. Similarly mathematical morphology is a tool for extracting image components that are useful in the representation and description of region shape, such as boundaries, skeletons and the convex hull. We also have morphological techniques for pre- or post processing such as morphological filtering, thinning, and skeletonization, pruning [4]. We present an algorithm that has high degree of accuracy on different type of Punjabi fonts and sizes. In this technique we use Morphological operator’s branch points, end points and thinning of binary images.

Thinning: Thinning means reducing binary objects or shapes in an image to strokes that are a single pixel wide.

Skeletonization: It is another process that reduce binary image object to a set of thin stroke that retain important information about the shape of the original objects.

2. Pre-Processing

Pre-processing is done to remove the noise and extra objects in the image so that only the character to be recognized remains in the image. There are several methods for pre-processing: by increasing the intensity level or applying various types of filtering etc. To remove the extra object, calculate the area of objects in the image and sort them in descending order. The largest area is the character area. Except that remove all objects in image based on area.

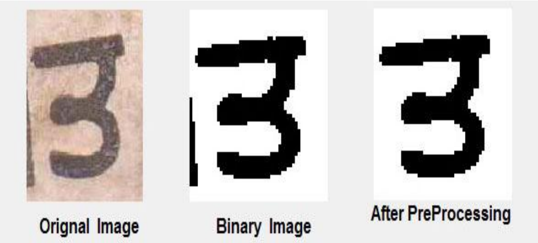


Figure 1. Pre-processing of Image

3. Feature Extraction

For the better generalization capability and low computational cost, we considered only three features of characters: holes, junctions and ends. The recognition is based on the number and position of these features. After pre-processing, the image is normalized to a dimension of 50x50, maintaining the aspect ratio.

3.1 Holes

The first feature we considered is the number of holes existing in the characters. In order to obtain an image in which every hole is represented as a point:

Fill the hole in input image by using following command

BWO = imfill(BWI,'holes');

Subtract the filled holed image from input image and shrink it, you will get the final image.

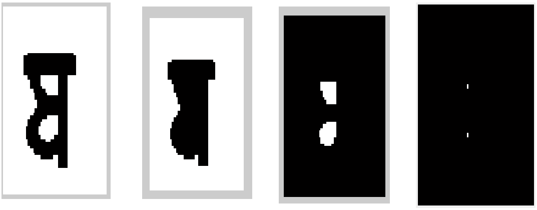


Figure 2. Hole detection

3.2 Junctions

Junction is also known as branch points. It's a point where two points meet. For example :

$$\begin{matrix}0&0&1&0&0\\1&1&1&1&1\\0&0&1&0&0\\0&0&1&0&0\end{matrix} \quad \text{becomes} \quad \begin{matrix}0&0&0&0&0\\0&0&1&0&0\\0&0&0&0&0\\0&0&0&0&0\end{matrix}$$

In matlab following command is used to extract the branch point

BWO = bwmorph(BWI,'branchpoints');

From Input Image BWI, and Return the Output Image BWO.
Count the number and position of junctions in the character, just by counting the number of black pixels.

3.2 Ends

An endpoint is a mark of termination or completion. Every character has a number of end points which play a significant role to recognize a character.

$$\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{matrix} \quad \text{becomes} \quad \begin{matrix}1&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{matrix}$$

In matlab following command is used to find the endpoint

BWO = bwmorph(BWI,'endpoints');

For Input image BWI and output Image BWO

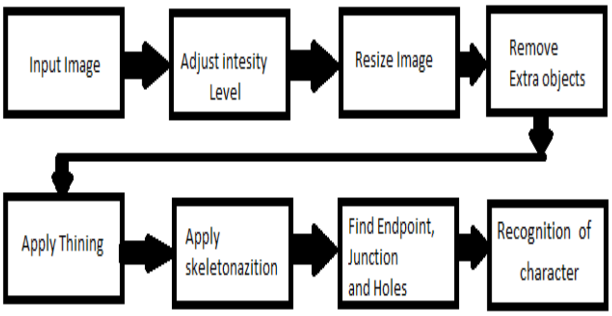


Figure: 3 Steps for the whole procedure

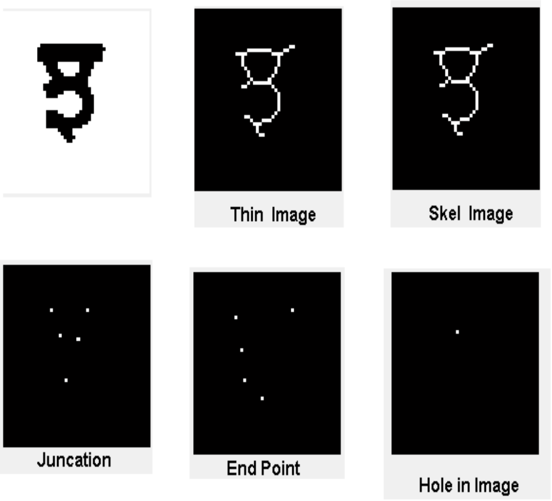


Figure: 4 Steps in fig: 3 applied to binary image containing character

4. Classification

First we performed the pre-processing of image , in pre-processing we adjusted intensity of image so that it get perfectly converted into binary image, then removed the extra area or any type of noise in the image, by just calculating the area of objects in the image . The largest area will represent character and all other areas are noise. So, by keeping the largest area we removed the areas representing noise.

Secondly, hole detection is done by filling the holes in input image and taking logical difference with original image and shrink it.

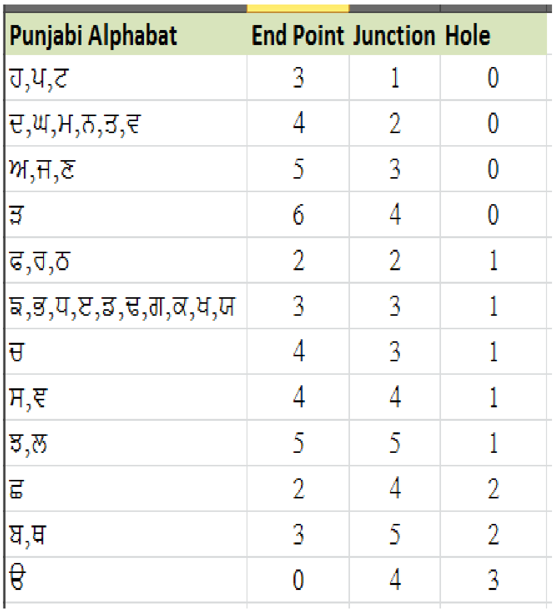
Then after thinning of image the **Skeletonization** applied to the image and the result will be used for both junction detection and for ends detection.

A first rough classification can be obtained just by counting the number of holes, junctions, and ends. For example, if the letter has three holes, we can immediately recognize it as ੳ, if it has two holes, and four junctions and two end point, it is ਛ. Nevertheless, the recognition of a letter by simply counting the number of holes, junction and ends is sometimes not possible: for example, if we count 0 holes, 1 junction, and 3 extremities, the letter could be ਹ ਪ ਟ. In this case, we can simply discriminate them just by looking at the position of the feature points too:

Divide the image into two equal parts upper half, and bottom half similarly left half and right half.

If the junction is at the right side and upper half, similarly if two end point is in right and one endpoint in left half and all three endpoint is lie in upper half ,and no hole, then the letter is ਹ.

The decision tree, for the first part is reproduced in Figure5.



Punjabi Alphabat	End Point	Junction	Hole
ਹ,ਪ,ਟ	3	1	0
ਦ,ਘ,ਮ,ਨ,ਤ,ਵ	4	2	0
ਅ,ਜ,ਣ	5	3	0
ੜ	6	4	0
ਫ,ਰ,ਠ	2	2	1
ਙ,ਭ,ਧ,ੲ,ਡ,ਢ,ਗ,ਕ,ਖ,ਯ	3	3	1
ਚ	4	3	1
ਸ,ਞ	4	4	1
ਝ,ਲ	5	5	1
ਛ	2	4	2
ਬ,ਥ	3	5	2
ੳ	0	4	3

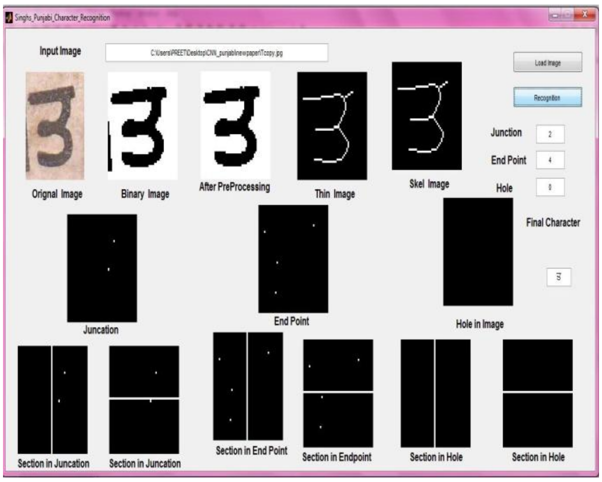
Figure 5: Decision Table

5. Experiment

Table 1. Recognition Rate for Different type of fonts

Font	Recognition Rate
AnmoLlipi	**100%**
Amrlipi	**100%**
GurbaniAkhar	**94.28%**
Asees	**97.14%**
AmrNeon	**91.42%**

We developed an application using matlab 2009 and conducted test on different Punjabi fonts and characters of different size. And result is shown below. Input to this application is any scanned image magazine, newspaper, old Punjabi record.



Input Image C:\Users\PREET\Desktop\CNN_punjabinewspaper\Tcopy.jpg

Junction 2

End Point 4

Hole 0

Final Character ਤ

Figure 6: Application to conduct Experiment

6. Result

The accuracy of this system is very good as shown in table; accuracy is down only in the cases where we need to divide the image into two or four equal parts. If any other languages which do not need to divide the image further equal parts for character recognition, this system will give the accuracy of **100%** .

7. Conclusion and Future Work

A method for recognition of Punjabi characters in machine printed documents is developed based on the morphological operators. The capabilities of this operator in detecting patterns with specific geometric properties in the image, is used to accomplish different essential tasks in a pattern recognition

process. This algorithm can be used for different and it can be extended to other languages.
We have implemented this algorithm on Punjabi characters; we will try this on complete Punjabi word. For example

□□□□ , ਹਜਰਤ

REFERENCES

[1] S. Kahan, T. Pavlidis, H. S. Baird, “On the recognition of printed characters of any font and size” IEEE Trans. Pattern Anal. Machine Intell., vol. 9, no. 2, pp. 274-288, March 1987.
[2] R.E. Howard, B. Boser, J.S. Denker, H.P. Graf, D.Henderson, W. Hubbard, L.D. Jackel, Y. LeCun, H.S. Baird: “Optical character recognition: a technology driver for neural networks” Circuits and Systems, 1990, IEEE International Symposium
[3] A Cellular Neural Network based character recognition system by Daniele Casali,Giovanni Costantini, Massimo Carota
[4] Book Digital image processing using Matlab by Rafael C. Gonzalez and Richard E. Woods
[5] M. Salman Jelodar, M.J. Fadaeieslam, N. Mozayani, M. Fazeli “A Persian OCR System using Morphological Operators” World Academy of Science, engineering and technology 4 2005.
[6] Seera, J, Image Analysis and Mathematical Morphology”,Acrdemic Press, New York,1982
[7] B. Timsari, Character recognition in typed Persian words: a morphological approach, M.S. Thesis, Isfahan Univ. Of Tech., Iran,1992
[8] J.W Smith and Z. Merali, “Optical character Recognition”, The British Library, Wetherby, West Yorkshire LS23 7BQ,UQ,1985