A Survey Paper on the Feature Extraction Module of Offline Handwriting Character Recognition
Abhishek Phukan¹, Mrinaljit Borah²

¹Sikkim Manipal Institute of Technology
Majitar, East Sikkim-737136
²Jorhat Engineering College
Garmur, Jorhat

ABSTRACT

Offline Character Recognition Is the process in image processing where a document is fed into the system in the form of an image for it to be recognized and classified. The feature extraction module is the basis for this paper and we give the different algorithms used in the feature extraction module after giving a brief overview of what a character recognition system is.

Keywords: character recognition, OCR, Feature Extraction Module

INTRODUCTION

Character recognition is the process of feeding characters into the system for it to be recognized and classified. The entire process can be classified as Online Character Recognition and Offline Character Recognition. We focus on the feature extraction module of the offline character recognition, also known as the OCR (Optical Character Recognition).

A. Online Character Recognition:

Online character recognition is the process of recognizing handwriting, recorded using a digitizer as a time sequence of the pen coordinates [1]. Digitizers are tablets that work either in a capacitive or in a resistive manner (pressure needs to be applied). The digitizers send in the coordinates at regular time intervals and the system senses the motion of the pen to recognize the characters. Online character recognition requires less pre-processing compared to Offline Character Recognition.

B. Offline Character Recognition:

Offline Character Recognition is the processing of feeding in the document into the system in the form of an image. Offline Character Recognition requires a much higher level of
pre-processing due to noise introduced by the scanning devices. The age of the paper or document also determines the amount of noise in the image. The stages in OCR include:

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post Processing

The next section of the paper focuses on the Feature Extraction Module. We first describe the Feature Extraction Process and then continue the discussion to the different works on the same that have been done.

[2] Survey on Feature Extraction of OCR

Feature Extraction is the process of extracting features from the segmented characters in order to recognize and classify the segmented characters in the classification phase. A single feature or multiple features can be extracted and combined to aid in the classification. Feature Extraction simplifies the amount of data required to recognize the character in the next phases. Features should be extracted in a way so that they can distinguish between different characters and for that reason, instead of taking a single feature, multiple features are extracted and combined together. In this section we will give an overview of the feature extraction methods followed by some of the works that have been implemented.

There are several methods or features that can extraction. Below given are some of the features.

A. Histogram Profile

Horizontal and vertical histogram profiles can be extracted as a feature of recognition. Horizontal projection can also be used to distinguish between different languages. For example, the Assamese language has a matra on the top whereas English does not. On extraction of the horizontal projection the matra forms a straight line and can be used to distinguish since in English there is no matra and hence no straight horizontal line will be created. [2]

B. Topological features

Topological features are another way of recognizing the characters. Topological features include dots, joints, loops and endpoints [2]. The number of loops for an example can be used to distinguish between the ‘g’ which has two loops and say ’q’ that has a single loop.

C. Crossing and Distances

The transition from the background pixel to foreground pixels along a vertical line forms the vertical crossing and similarly the transition in the horizontal line forms the horizontal crossing. Distances are measured from the boundaries of the image to the point where the first pixel is detected [5]. The distance can be taken from all 4 boundaries or even in a diagonal feature so that it gives us 8 distances.
D. Structural Features

Structural features are also an efficient way of distinguishing the characters. Structural features include number of line segments (Horizontal and Vertical), number of strokes and say number of curves [5]. Any structure of a character can be taken to distinguish it from another character. For example, ‘E’ has 4 line segments (1 Vertical and 3 Horizontal) and ‘H’ has 3 line segments (2 Vertical and 1 Horizontal) and hence this can be used to distinguish between them.

The Fourier transformation is another characteristic that can be taken as a feature.

E. Chain Code of the contour

Chain code can be used to find the direction of the contour which can be then used as a feature [6]. A point is taken as the reference point (say the midpoint) and then the nearby pixels are represented from 0-7 depending on the position of the pixel with relative to the reference point. The whole direction of the contour can be calculated this way.

F. Zoning

Zoning is done to extract the local characteristics rather than the global characteristics. Taking local characteristics from an image increases the accuracy of the character classification.

These are some of the common features that can be extracted to classify a character. In the next section we give an overview of the different works in which features have been extracted and also new features have been taken for a better efficiency. Instead of taking these features one by one, they can also be combined in order to give a greater efficiency and enhance character classification. In the following section we discuss some of the works done in feature extraction for offline handwriting character recognition.

S. Saha, N. Paul, S. K. Das and S. Kundu in their paper [7] have used a 40 point Feature Extraction. The PreProcessing phase comprises of cropping, grayScaling and Binarization of the image. They in their work have first zoned the image into 16 zones in a regular fashion. In the consequent stages the image is further again zoned diagonally from all 4 corners. In this fashion 28 zones are created. After this the innermost 4 cells are taken as zone 29 and expanded to form zones till 31. In a similar fashion the rest of the zones i.e. from zones 32 to 40 are created. The zones are overlapped and features are extracted from these zones. An accuracy in the range of 70% to 100% is achieved depending on the character is achieved. They used the English Alphabets for their work. N.Arica and Fatos T. Yarman-Vural have shown in their paper [1] the different feature extraction methods. They have classified the entire process into Global Transformation and Series Expansion, Statistical Representation and Geometrical and topological representation. Fourier transformation, Gabor transformation, wavelet, moments and Karhunen-Loeve Expansion are taken in the global transformation and series expansion. For statistical and geometrical features, they have considered zoning, crossing and distances and Projections. The explanation for these features have been given in the previous section.
Calculations of the maxima and minima on the exterior and interior contours, reference lines, ascenders and descenders are taken as the topological features. G. Y. Tawde and Mrs. J. M. Kundargi in their paper [8] have taken Indian Scripts as the basis for feature Extraction. After the preProcessing and the Segmentation phase, the segmented image is first divided into zones (both overlapping and non-overlapping). They have from the zones then extracted features such as density and contour direction. Next the projection and profile of the characters are extracted. The profile counts the distance between the bounding box and the edge of the character. Next the crossing and distances and the global transformations and moments of the image are calculated. These features are combined in the classification phase to recognize the characters. R. Tokasl, and A. Bhadu in their paper [5] have given a comparative analysis on the different feature extraction techniques. They have given the merits and demerits of all the methods analysed. V. P. Agnihotri in his paper [9] have used Devanagiri script for character recognition. The image is divided into 54 zones. Diagonal Feature Extraction scheme is used in his work. 19 features are extracted from each zone. The features are then made into a chromosome string for classification (Genetic Algorithm). In his work, V. P. Agnihotri has achieved a match of 85.78% and 13.35% mismatch. O. Surinra, L. Schomaker and M. Wiering in their paper [10] have used a HotSpot Feature Extraction method. The distance between the black pixels and hotspots are used as a feature. The direction of hotspot is defined using the chain Code and the distance to the last HotSpot is calculated. By this method they have achieved an accuracy of about 87.8%. J. Pradeep, E. Srinivasan and S. Himavathi in their paper [11] have used diagonal based feature extraction for recognition of offline handwriting recognition. The image is divided into 54 zones and each zone is traversed in a diagonal fashion to extract 19 sub features for each zone. The 19sub features are averaged to form a single feature and hence for each character there are 54 features. They have achieved an accuracy of 97.8% for 54 features and 98.5% for 69 features. Dinesh Dileep in his paper [12] has shown the use of character geometry as the feature to be used for the classification phase. The features in his work include geometrical features such as the number of horizontal and vertical lines, number and length of the right diagonal and left diagonal lines. The image is first zoned into 9 zones and these features are extracted then from these individual zones. The algorithm was tested on 130 characters out of which 6 were detected erroneously

[3] Conclusion

In this paper we have given a brief overview of the OCR process, also known as the offline handwriting recognition. In this paper we have focused mainly on the Feature Extraction Module of the process. The feature extraction is the most important part of the entire process where the features of the characters are extracted so that they can be used to classify characters later on. For an efficient recognition we need to take a number of features for the same alphabet and also the same features must be extracted from different handwriting. This way an average can be found out and the features of the test character can then be compared with the database. The larger the database, the more will be the accuracy. After giving a few common features that
can be extracted to help in the classification phase, we have given an overview of the works done in the same by different people and also the level of accuracy that they have obtained.

REFERENCES


[10] Olarik Surinta, Lambert Schomaker, Marco Wiering, “Handwritten Character Classification Using the HotSpot Feature Extraction Technique”


[12] Dinesh Dileep, “A Feature Extraction Technique based on Character Geometry For Character Recognition”.

Abhishek Phukan and Mrinaljit Borah