Structural and Statistical Feature Extraction Methods for Character and Digit Recognition

Purna Vithlani
Research Scholar,
Department of Computer Science,
Saurashtra University, Rajkot

C.K.Kumbharana, Ph.D
Head,
Department of Computer Science,
Saurashtra University, Rajkot

ABSTRACT
Character recognition is the process of converting an image or PDF file into editable and searchable text file. Feature Extraction is the heart of any character recognition system. In the proposed paper, structural and statistical features of isolated English capital characters and digits are presented. New features like end point existence is zone and zone with zero foreground pixel value are presented in this paper. Features like zoning, number of endpoints, end point existence in zone, zone with zero foreground pixel value, number of horizontal line and number of vertical line are extracted from character.

General Terms
Character Recognition, Structural Features, Statistical Features

Keywords
Zoning, Preprocessing, Character Segmentation, Feature Extraction, Classification

1. INTRODUCTION
Character recognition is the process of converting an image file to text file. In any character recognition system, a scanned image of a document is taken as an input then preprocessing, character segmentation, feature extraction and classification steps are performed on it.

Pre-processing is necessary to modify the raw data to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor. Pre-processing is the preliminary step which transforms the data into a format that will be more easily and effectively processed. Character segmentation is the process of separating characters from an image. Feature extraction is the process of identifying unique features of characters. Classification stage is the decision making part of a recognition system and it uses the features extracted in the previous stage. In the proposed paper, feature extraction step is explained in detail. Figure 1 represents character recognition process.

2. FEATURE EXTRACTION METHODS
The heart of any character recognition system is the formation of feature vector to be used in the recognition stage. Feature extraction can be considered as finding a set of features that define the shape of the underlying character as precisely and

Uniquely as possible. The term feature selection refers to methods that select the best subset of the input feature set. There are two feature extraction methods - Structural method and Statistical method.

2.1 Structural Method
Structural method identifies structural features of a character. Structural features are based on topological and geometric properties of the character [1]. Examples of structural features are number of horizontal lines or vertical lines, number of endpoints, number of cross points, horizontal curves at top or bottom etc. This type of representation may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object.

2.2 Statistical Method
Statistical method identifies statistical features of a character. The statistical features are derived from the statistical distributions of pixels [2]. These features can be easily detected as compared to structural features. Statistical features are not affected too much by noise or distortions as compared to statistical features. A number of techniques are used for statistical feature extraction; some of these are: zoning, projection histograms, crossings and distances, n-tuples.

In the proposed paper, Zoning, Number of endpoints, End point existence in zone, Zone with zero foreground pixel value, Number of horizontal line and Number of vertical line are extracted from character.

1. Zoning
Zoning based feature extraction is one of the most popular methods. The character image is divided into predefined number of zones and a feature is computed from each of these zones. The character image is divided into several overlapping or non-overlapping zones. A character is usually divided into zones of predefined size. These predefined sizes are typically of the order 2×2, 3×3, 4×4 etc. In Figure 2, Character A is divided into 3×3 zones.

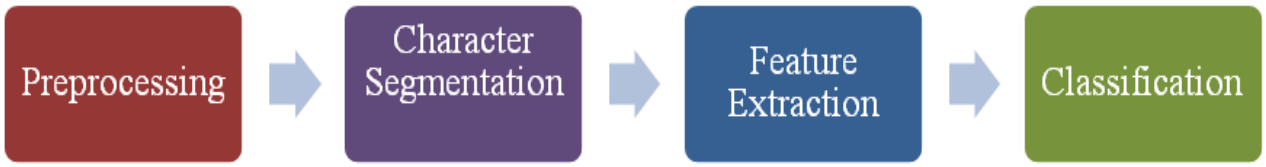

Figure 1 Character Recognition Process

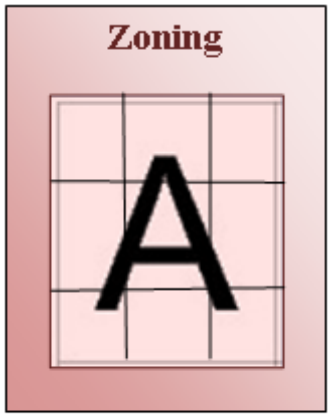

Figure 2 Zoning of Character A

2. Number of Endpoints

Endpoint defines starting and ending point of a character. As shown in Figure 3, Character Y has 3 endpoints.

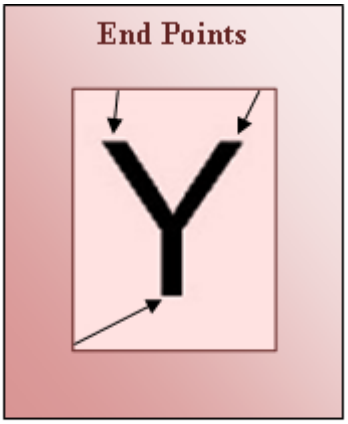

Figure 3 Character Y having 3 Endpoint

3. End Point Existence in Zone

After finding numbers of endpoints, check in which zone endpoint exists. As shown in Figure 4, Y has 3 endpoints and endpoint existences are in Zone 1, Zone 3 and Zone 7.

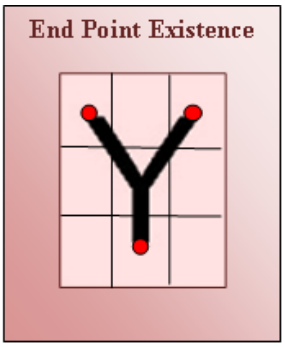

Figure 4 Endpoint Existence in Zone 1, Zone 3 and zone 8

4. Number of Vertical Lines

Vertical Line feature describes that character contains vertical line. In English language vertical lines are always connected with some other part of the character and are not independent. As shown in Figure 5, character H contains 2 vertical lines.

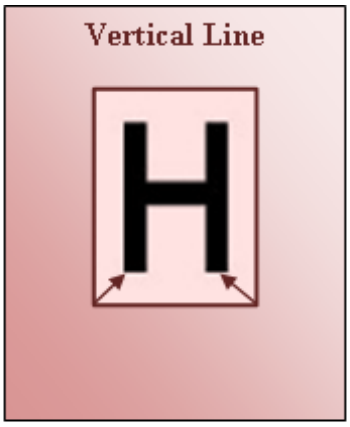

Figure 5 Character H having 2 Vertical Line

5. Number of Horizontal Line

Horizontal Line feature describes that character contains horizontal line. In English language horizontal lines are always connected with some other part of the character and are not independent. As shown in Figure 6, character E has 3 horizontal lines.

Figure 6 Character E having 3 Horizontal Line

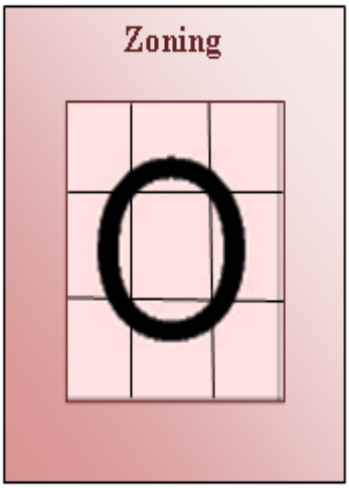

Figure 7 Zoning of Character O

6. Zone with Zero Foreground Pixel Value

For this feature, thresholding (black and white) image is used. Image is divided into 9 zones and Count the number of foreground (black) pixels in each zone and find out the zone with zero foreground pixel value. As shown in Figure 7 in character O, zone 5 have zero foreground pixel value.

3. ANALYSIS OF ENGLISH CHARACTERS FOR STRUCTURAL AND STATISTICAL FEATURES

Structural and statistical feature analysis of English capital alphabets is presented in Table 1.

Table 1 Structural and Statistical Features of Characters

Character	End Point	End Point Existence	Vertical Line	Horizontal Line
A	2	Zone 7, Zone 9	0	1
B	0	-	1	0
C	2	Zone 3, Zone 9	0	0
D	0	-	1	0
E	3	Zone 3, Zone 6, Zone 9	1	3
F	3	Zone 3, Zone 6, Zone 7	1	2
G	3	Zone 2, Zone 6, Zone 9	0	1
H	4	Zone 1, Zone 3, Zone 7, Zone 9	2	1
I	4	Zone 1, Zone 3, Zone 7, Zone 9	1	2
J	3	Zone 1, Zone 3, Zone 7	1	1
K	4	Zone 1, Zone 3, Zone 7, Zone 9	1	0
L	2	Zone 1, Zone 9	1	1
M	2	Zone 7, Zone 9	2	0
N	2	Zone 3, Zone 7	2	0
O	0	-	0	0

P	1	Zone 7	1	0
Q	2	Zone 5, Zone 9	0	0
R	2	Zone 7, Zone 9	1	0
S	2	Zone 3, Zone 7	0	0
T	3	Zone 1, Zone 3, Zone 8	1	1
U	2	Zone 1, Zone 3	2	0
V	2	Zone 1, Zone 3	0	0
W	2	Zone 1, Zone 3	0	0
X	4	Zone 1, Zone 3, Zone 7, Zone 9	0	0
Y	3	Zone 1, Zone 3, Zone 7	0	0
Z	2	Zone 1, Zone 9	0	2

As shown in Table 1, Characters B, D and V, W have similar features. To distinguish these characters, Researchers have identified a new feature - Zone with zero foreground pixel value. To distinguish B and D, divide a character into 9 zones and count the number of foreground pixel values of Zone 5. If foreground pixel values of zone 5 is 0 then it is D else it is B. To distinguish between V and W, divide a character into 9 zones and count the Number of foreground pixel values of Zone 2. If foreground pixel values of zone 2 is 0 then it is V else it is W.

4. ANALYSIS OF ENGLISH DIGITS FOR STRUCTURAL AND STATISTICAL FEATURES

Structural and statistical feature analysis of digit is presented in Table 2.

Table 2 Structural and Statistical Features of Digits

Digit	End Point	End Point Existence	Vertical Line	Horizontal Line
1	3	Zone 1, Zone 7, Zone 9	1	1
2	2	Zone 1, Zone 9	0	1
3	2	Zone 1, Zone 7	0	1
4	2	Zone 6, Zone 8	1	1
5	2	Zone 3, Zone 7	1	1
6	1	Zone 2	0	0
7	2	Zone 1, Zone 8	0	1
8	0	-	0	0
9	1	Zone 8	1	0
0	0	-	0	0

As shown in Table 2, Digits 8 and 0 have similar features. To distinguish these digits, Zone with zero foreground pixel value is identified. To distinguish 8 and 0, divide a digit into 9 zones and count the number of foreground pixel values of Zone 5. If foreground pixel values of zone 5 is 0 then it is 0 else it is 8.

5. CONCLUSION

In this paper, structural and statistical features of English capital characters and digits are described and analysis of them is presented. Researchers have presented new features like End point Existence in zone and Zone with zero foreground pixel value.

6. REFERENCES

[1] Giorgos Vamvakas, "Optical Character Recognition for Handwritten Characters".

[2] Sumedha B. Hallale, Geeta D. Salunke, "Twelve Directional Feature Extraction for Handwritten English Character Recognition", International Journal of Recent Technology and Engineering (IJRTE), Vol.-2, Issue-2, May 2013.