ISSN:0975-9646

Zone Based and Morphological Based Feature Extraction Techniques for English Alphabets

Saleem Pasha[#1], Ajay Kumar[*2], Syed Younus Hussain[#3], Amritansh[*4], Mohammed Shafi Ul Umam[#5]

[#]*Information Science and Engineering Department, P.E.S. College of Engineering, Mandya, Karnataka, India*

[*]*Information Science and Engineering Department, P.E.S. College of Engineering, Mandya, Karnataka, India*

***Abstract*— Developing an efficient and robust Optical Character Recognition (OCR) system is the most interesting and challenging area in the field of Image processing. OCR is a process that provides a full alphanumeric recognition of printed or handwritten characters. This paper focuses on developing an OCR system, which consist of phases such as pre-processing, feature extraction and classification. Two feature extraction techniques such as zone based and morphological operation based techniques are proposed to recognize the alphabets present in an image using nearest neighbor classifier. An attempt is made to compare this two feature extraction techniques and their accuracies are recorded.**

Keywords— **Optical Character Recognition, zone based, morphological operation, nearest neighbor.**

I. INTRODUCTION

During the past three decades, numerous methods have been proposed for machine recognition of handwritten characters, especially for languages such as English, Japanese, Chinese, so on. Optical Character Recognition (OCR) is a type of machine recognition system, which can be used for commercial purpose. OCR system is used to recognize the text present in the scanned image, which are a most fascinating and challenging areas of image processing with various practical applications. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine.

Character recognition can be classified into two categories: offline recognition and online recognition. Offline character recognition include recognition of machine printed, hand printed and handwritten characters, from the scanned images. On-line handwritten characters are obtained by using cameras or by writing the characters on a sensitive surface.

In this paper, offline recognition (OCR) system is developed for recognizing the English alphabets. The reason for using the English language as a case study is due to its global usage. The OCR system usually consist of the following phases such as pre-processing, feature extraction and classification. The input image is passed through suitable pre-processing, feature extraction and classification techniques to obtain the recognized alphabet. In this paper, zone based and morphological operation based feature extraction techniques are proposed and recognition of English alphabets is done using nearest neighbor classifier.

II. PREVIOUS WORK

Previous works relevant to our work is highlighted. Ravi Kumar et al have proposed the recognition of English characters by codes generate using neighbour identification [1]. They have used simple chain codes that are generated by locating the pixels and identifying the orientation of neighbour pixels. Matching is done using Longest Common Subsequence Algorithm. They have tested for different fonts and with respect to the fonts accuracy is tabulated.

Prerna Kakkar et al have developed a novel approach to recognition of English characters using Artificial Neural Network, which provides a very high recognition rate for all English alphabets in the presence of noise [2].

Gaganjot Kaur et al have proposed Artificial Intelligent System for character recognition using Levenberg-Marquardt Algorithm [3]. Printed characters are considered and obtained an overall accuracy of 93%.

Rakesh Kumar Mandal et al have developed hand written English character recognition using Row wise Segmentation Technique (RST) [4]. This work is an approach to develop a method to get the optimized results using the easily available resources. As they have considered handwritten characters, they have obtained an overall accuracy of 80%.

Yusuf Perwej et al have proposed neural network based handwritten English alphabet recognition [5]. In this system, each English alphabet is represented by binary values that are used as input to a simple feature extraction system and the output is fed to the neural network system. They have obtained an average accuracy of 82.5%.

Ranpreet Kaur et al have developed a hybrid neural approach for character recognition system [6]. This work describes the process of character recognition using the hybrid algorithm of Back Propagation and Genetic Algorithm for the recognition of uppercase alphabets and obtained an accuracy of 91.1% .

Purna Vithlani et al have proposed a study of optical character patterns identified by the different OCR Algorithms [7]. A study is carried out using different classifiers such as Template Matching Algorithm, statistical Algorithm, Structural Algorithm, Neural Network Algorithm and Support Vector Machine Algorithm. Comparison with respect to efficiency of these classifiers is tabulated.

In this paper, features are extracted using two techniques such as zone based and morphological operation based

techniques and recognition of English alphabets is performed using nearest neighbour classifier.

III. PRE-PROCESSING

Pre-Processing can be defined as cleaning the document image and making it appropriate for better feature extraction. Major methods such as binarization, crop to edge, noise removal, normalization, and thinning are considered under pre-processing.

A. Binarization

The task of binarization is to extract the foreground from the background. Given a threshold 'T' between 0 and 255, replace all the pixels with gray level lower than or equal to T with black (0) and the rest with white (1). The decision of the appropriate threshold value is chosen globally or locally. This approach is called thresholding, in which gray-scale image is converted into binary image.

B. Noise Removal

Noise is defined as any degradation in the image due to external disturbance. The Noise introduced by the optical scanning devices in the input leads to poor system performance. These imperfections must be removed prior to character recognition. Noise can be of different types such as Gaussian noise, Gamma noise, Rayleigh noise, Median filtering, Exponential noise, Uniform noise, Salt and pepper noise, Periodic noise etc. in this paper, median filtering is used for noise removal. The reason for using median filter is it is more effective, when the goal is to simultaneously reduce noise and preserve edges.

C. Normalization

Normalization is the process of converting the random sized image into standard sized image. This size normalization avoids inter class variation among characters. Before applying normalization, extra white spaces (background space) present in the image is removed. Finally, the given input image is normalized to a standard resolution of 128×128.

D. Thinning

Thinning is an image preprocessing operation performed to make the image crisper by reducing the binary valued image regions to lines that approximate the skeletons of the region.

IV. FEATURE EXTRACTION

Feature extraction is the important phase of OCR system, which is used to find the set of parameters that define the shape of a character precisely and uniquely. Features are a set of numbers that capture the salient characteristics of the pre-processed image. There is different feature extraction methods proposed for character recognition. Suitable features are extracted to perform feature extraction.

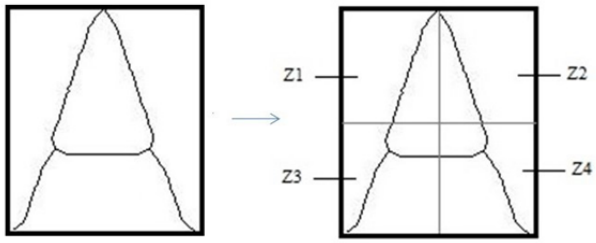
In this paper, we are extracting different features using two feature extraction techniques such as,

1) Zone based technique
2) Morphological operation based technique

A. Zone Based Technique

In zone based feature extraction, the input image is divided into different zones as required. These zones can be interpreted as parts for understanding purpose. The divided zone is then taken as one sub image and suitable features are extracted. This process is repeated for all the remaining zones.

As shown in figure 1, the pre-processed image is passed through zone based technique, where the pre-processed image is divided into four zones. From each zone, a set of five features are extracted such as Zone Area, horizontal features, vertical features, right diagonal features and left diagonal features. Same five features are extracted from remaining zones. Finally, a total of twenty features are extracted from the entire image. The reason for using these features is due to the presence of different segments such as horizontal line, vertical line, left diagonal, right diagonal, difference in area in every English alphabet. These five features are explained below.



Pre-processed Image 'A' Image 'A' after Zoning

Fig. 1 Zone Based Technique

These five features are explained below.

1) Horizontal line: It is a 3X3 mask in which all the elements in the second row are 1 and rest of the elements are 0. Horizontal Line (HL) is calculated using the equation (1).

HL = number of horizontal components × length of all the horizontal components. (1)

2) Vertical Line: It is a 3X3 mask in which all the elements in the second column are 1 and rest of the elements are 0. Vertical Line (VL) is calculated using the equation (2).

VL = number of vertical components × length of all the vertical components. (2)

3) Left diagonal: It is a 3X3 mask in which all the elements in the left diagonal of the mask are 1 and rest of the elements are 0. Left Diagonal (LD) is calculated using the equation (3).

LD = number of left diagonal components × length of all the left diagonal components. (3)

4) Right diagonal: It is a 3X3 mask in which all the elements in the right diagonal of the mask are 1 and rest of the elements are 0. Right Diagonal (RD) is calculated using the equation (4).

RD = number of right diagonal components × length of all the right diagonal components. (4)

5) Zone Area: It is defined as the ratio of number of black pixels in each zone to the total number of black pixels in the whole image.

In addition to these five features, two more features are extracted from the whole image. These two additional features are Euler number and number of blobs.

6) Euler number: It specifies the number of objects in the region minus the number of holes in those objects.

7) Blobs: Blobs are the area enclosed within the closed loops of an object in an image which can be filled with colours. It is one of the features extracted in the feature extraction phase to count the number of enclosed loops present in the object of an image.

Hence, we obtain 20 features from zone based technique and 2 features such as Euler number, Blobs from the whole image. Finally, we obtain 22 features from zone based technique and store this 22 features in a feature vector.

B. Morphological Operation Based Technique

Morphology is a broad set of image processing operations that process images based on shapes. The most basic morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. In this paper, we are extracting features using two morphological operations such as Dilation and Hit-or-Miss Transform.

1) Dilation: Dilation allows objects to expand, thus potentially filling in small holes and connecting disjoint objects. The dilation process is performed by laying the structuring element on the image and sliding it across the image.

2) Hit-or-Miss Transform: The hit-and-miss transform is a general binary morphological operation that can be used to look for particular patterns of foreground and background pixels in an image. The goal of the "hit-miss" operation is to find pixels **x**, for which B^1_x is in A ("hit") and where no pixel in B^2_x is in A ("miss"), thus B^2_x is in A^c.

Steps in Morphological Operation Based Technique:

1) Input image is converted to pre-processed image, say 'A'.
2) Structuring element is defined, say 'SE'.
3) Dilation operation is applied on the pre-processed image 'A' using the structuring element 'SE' and we obtain a dilated image, say 'B'.
4) Difference between dilated image and pre-processed image is obtained and saved in 'C'.
5) Finally, we apply Hit-or-Miss Transform with the following inputs, such as testing image or training image say 'Z', pre-processed image 'A' and difference image 'C'.
6) Depending on the non zero value of hit-or-miss transform, corresponding alphabet is displayed.
7) The above six steps are repeated for all the English alphabets.

In this technique, we are finally recognizing the alphabet using only morphological operations.

V. CLASSIFICATION

Recognition of the English alphabets is done using classification techniques. We have proposed two feature extraction techniques in feature extraction phase. Zone based technique requires classification phase, but morphological operation based technique does not require classification phase.

Recognition of alphabets is carried out using nearest neighbor classifier for the image containing features using zone based technique. Nearest neighbor classifier is an instance-based learning or lazy learning, where the function is only approximated locally and all computation is deferred until classification. In this classifier, we use a K value which is a positive integer. In this paper, K-nearest neighbor classifier is used for the recognition of English alphabets and obtained a better accuracy for the value of K=7.

VI. ALGORITHM FOR PROPOSED OCR SYSTEM

Algorithm Begins

Step 1: Change RGB Image to Binary Image.
Step 2: Remove the noise.
Step 3: Normalize the Image to Standard size.
Step 4: Thinning of image is performed and finally, a pre-processed image is ready.
Step 5: Apply two feature extraction techniques such as zone based technique and morphological operation based technique.

1) For zone based technique, the pre-processed image is divided into four different zones. From each zone, five different features such as vertical line, horizontal line, left diagonal, right diagonal and zone area are extracted. Finally, from four zones, a total of 20 features are extracted. In addition to these 20 features, two additional features are extracted from the entire image. They are blob and Euler number. Hence, 22 features are extracted from zone based technique and stored in a vector.
2) For morphological operation based technique, dilation and hit-or-miss transform is used and recognition is done.

Step 6: Apply nearest neighbor classifier for recognition of English alphabets for 1st technique of Step 5.

Algorithm Ends

VII. EXPERIMENTAL RESULTS AND DISCUSSION

OCR is an active topic in image processing and pattern recognition. The printed English alphabets are considered with different fonts. The data set has been created for the experimentation. At present, we have considered clear printed alphabets. The proposed model is implemented using Matlab in Windows 7 platform. For extracting the features, a total of 1040 training samples are used. For recognition purpose, a total of 260 testing samples are used. The recognition of English alphabets is achieved using K-nearest neighbor classifier.

TABLE I
RECOGNITION RATE

Alphabets	No. of training samples	No. of testing samples	Accuracy using Zone based (%)	Accuracy Using Morphological Operation (%)
A	40	10	90	80
B	40	10	100	90
C	40	10	80	90
D	40	10	80	80
E	40	10	100	80
F	40	10	90	90
G	40	10	90	70
H	40	10	90	80
I	40	10	90	80
J	40	10	90	70
K	40	10	80	80
L	40	10	90	90
M	40	10	80	80
N	40	10	90	80
O	40	10	100	90
P	40	10	90	80
Q	40	10	80	80
R	40	10	90	90
S	40	10	90	80
T	40	10	100	90
U	40	10	90	90
V	40	10	90	70
W	40	10	90	80
X	40	10	90	90
Y	40	10	100	80
Z	40	10	100	80
Average Accuracy			**90.38**	**82.38**

The performance of K–nearest neighbor classifier was better when the value of K=7 is used. Recognition rate for different alphabets is given in Table 1. Figure 2 shows the accuracy graphically for all the alphabets using the proposed two feature extraction techniques.

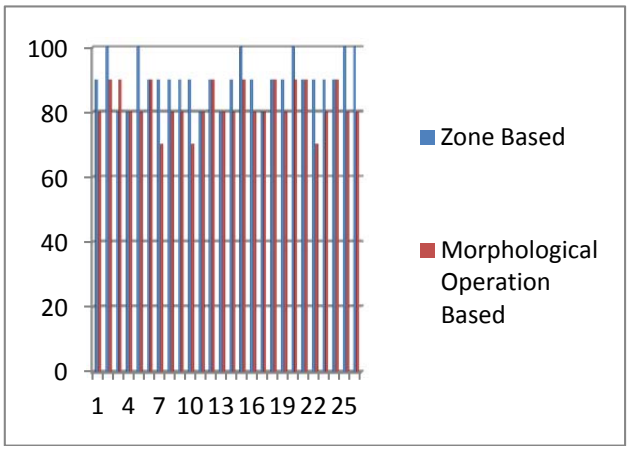


Fig. 2 Accuracy Shown Graphically

The accuracy obtained using zone based technique is 90.38% and accuracy obtained using morphological operation based technique is 82.38. The reason for reduction in accuracy for morphological operation based technique may be due to usage of only two morphological operations.

VIII. CONCLUSIONS

In this paper, an attempt is made to develop an Optical Character Recognition (OCR) system. Feature extraction is implemented by proposing two techniques such as zone based technique and morphological operation based technique. Nearest neighbor classifier is used for recognition. An attempt is made to compare this two feature extraction techniques and it is found that zone based technique (90.38%) is better than morphological operation based technique (82.38%).

In future, we will consider printed English alphabets with more complexity and extend our work on handwritten English alphabets.

REFERENCES

1. Ravi Kumar, Anurag Anand and Nikunj Sharma, "Recognition of English characters by codes generated using neighbour identification," International Journal of Application or Innovation in Engineering and Management, Volume 2, Issue 4, April 2013, pp. 466-470.
2. Prerna Kakkar and Umesh Dutta, "A novel approach to recognition of English characters using artificial neural network," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 6, June 2014, pp. 10238-10245.
3. Gaganjot Kaur and Monika Aggarwal, "Artificial intelligent system for character recognition using Levenberg-Marquardt algorithm," International Journal of Advanced Research in Computer Science and Software Engineering, research paper, ISSN: 2277 128X, pp. 220-230.
4. Rakesh Kumar Mandal and N R Manna, "Hand written English character recognition using Rowwise Segmentation Technique (RST)," International Journal of Computer Applications, 2011, pp. 5-9.
5. Yusuf Perwej and Ashish Chaturvedi, "Neural networks for handwritten English alphabet recognition," International Journal of Computer Applications (0975 – 8887) Volume 20– No.7, April 2011, pp. 1-5.
6. Ranpreet Kaur and Baljit Singh, "A hybrid neural approach for character recognition system," International Journal of Computer Science and Information Technologies, Vol. 2 (2) , 2011, pp. 721-726.
7. Purna Vithlani and C. K. Kumbharana, "A Study of optical character patterns identified by the different OCR algorithms," International Journal of Scientific and Research Publications, Volume 5, Issue 3, March 2015, ISSN 2250-3153, pp. 1-5.