# Geometric Feature Points Based Optical Character Recognition

M. Usman Akram*, Zabeel Bashir †, Anam Tariq § and Shoab A Khan‡
Department of Computer Engineering
College of Electrical & Mechanical Engineering
National University of Sciences & Technology
Pakistan.
Email: usmakram@gmail.com*, zabeel_bashir89@hotmail.com†, anam.tariq86@gmail.com §, kshoab@yahoo.com‡

***Abstract*—Optical character recognition is an application of pattern recognition which automatically detects and recognizes the optical characters with out human intervention. All the characters are basically made up from three geometric entities, i.e. corners, endings and bifurcations which can be used to identify different characters. In this paper, we present a method for optical character recognition based on basic geometric features. The method uses a crossing number method to extract features from thinned character. The feature vector for each character consists of number of corners, endings and bifurcations. The classification stage recognizes a character by using a simple rule based method. The proposed system is tested using different samples for each character and the results show the validity of the proposed algorithm.**

## I. Introduction

Optical character recognition(OCR) is an interesting topic in pattern recognition [1]. In the field of pattern recognition, OCR is just like 3D object recognition and image retrieval systems [1]. The aim of OCR is to detect characters using computer programs and without any input from human. OCR has its applications in a number of fields such as postal code identification, automated guided vehicles, digital libraries, object identification, processing of different receipts and personal digital assistants [2].

An automated system for OCR consists of image acquisition, preprocessing, feature extraction, similarity measure between the extracted features and the features stored in database and finally character recognition based on similarity score. A number of OCR algorithms have been presented to improve the accuracy of character recognition. Most of these algorithms include character matching using templates, different image shapes, geometric properties and invariant techniques based on shapes [3]. OCR has its roots back in 1929 when Tausheck obtained a patent named "Reading Machine" in Germany [4]. The base of this machine was template matching which is still used in some applications. Some of studies have used algorithms consisting of locating and enhancing the character areas [5]-[8]. Other than these, shape based algorithms are of most importance. The shape based algorithms are invariant to translation, rotation and scaling. They are broadly divided in two categories, i.e. boundary based and shape based algorithms [9]. Plamondon et al [10] and Pal et al [11] proposed methods for off-line hand printed characters recognition. The intra class variation problem has been solved by them using different styles of writing. A natural language processing based method was used in [12]. They used a Markov chain framework for parsing images. In most of cases, the feature extraction and classifiers need extensive efforts while we present a method which uses geometric features and simple rule based classifier.

This paper contains four sections. Section 2 explains the proposed system in detail. It contains preprocessing, feature extraction and classification. The evaluation results for proposed system are stated in section 3. Last section summarizes the system.

## II. Proposed System

The proposed system is divided into two stages, i.e. preprocessing and feature extraction. The preprocessing stage removes the noise and converts the width of character to single pixel. Second stage is feature extraction which applies a modified crossing number method to extract geometric features from the thinned character. Each character can be differentiated based on number of corners, endings and bifurcations, it has. In classification, the systems uses a simple rule based method to identify the input character. Figure 1 shows the flow diagram and different stages of proposed system.

### A. Preprocessing

The purpose of preprocessing stage is to make the image suitable for feature extraction algorithm. The input image may contain noisy areas or any background pixels which should be removed to improve the accuracy of feature extraction and classification. The proposed preprocessing stage consists of background estimation, noise removal, adaptive thresholding and morphological thinning. A simple thresholding technique proves to be useless because of the streaked nature of the input character images. The presence of noise in image requires more dynamic techniques for accurate detection of characters. A good preprocessing method should be insensitive to image contrast and should detect smudged or noisy regions. So, we present a robust preprocessing method to cater variable background and noisy areas. The algorithm divides the image $I$ into non-overlapping blocks of size $k \times k$ and applies histogram

978-1-4799-1125-7/13/$31.00 ©2013 IEEE

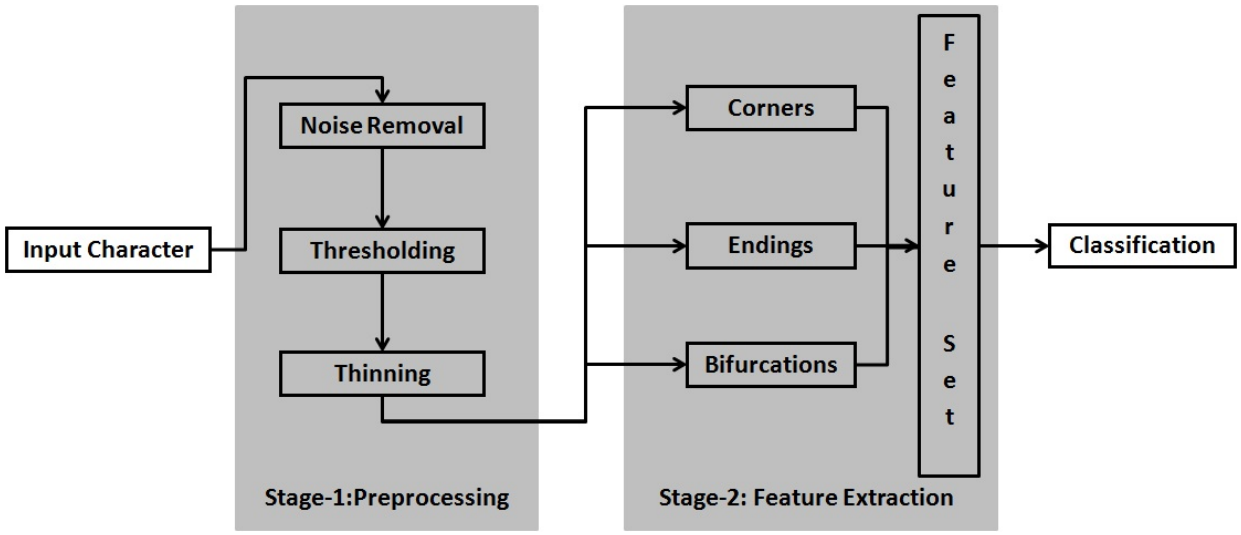


Fig. 1. Flow diagram for proposed system

equalization to enhance the contrast between background and foreground. An adaptive median filer is applied to remove the noise if present. In order to estimate the background, the algorithm computes the gradients $\partial_x$ and $\partial_y$ for input image. The mean and standard deviation values for $x$ and $y$ components of gradient are computed using equations 1, 2, 3 and 4 respectively [13].

$$M_x = \frac{1}{k^2} \sum_{i=-k/2}^{k/2} \sum_{j=-k/2}^{k/2} \partial_x(i,j) \tag{1}$$

$$M_y = \frac{1}{k^2} \sum_{i=-k/2}^{k/2} \sum_{j=-k/2}^{k/2} \partial_y(i,j) \tag{2}$$

$$std_x = \sqrt{\frac{1}{k^2} \sum_{i=-k/2}^{k/2} \sum_{j=-k/2}^{k/2} (\partial_x(i,j) - M_x(I))^2} \tag{3}$$

$$std_y = \sqrt{\frac{1}{k^2} \sum_{i=-k/2}^{k/2} \sum_{j=-k/2}^{k/2} (\partial_y(i,j) - M_y(I))^2} \tag{4}$$

Finally the system computes the gradient deviation using equation 5.

$$grddev = std_x + std_y \tag{5}$$

In order to separate the character from background and convert the image into binary, we apply adaptive threshold [13]. The binary character contains the letters with variable width which make feature extraction difficult. In order to bring all characters to single pixel width, we apply morphological thinning operation [14]. Figure 2 shows the binary and thinned character.

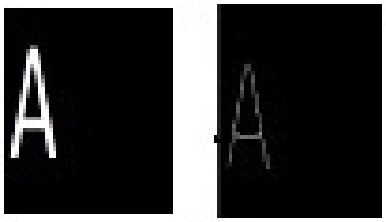

Fig. 2. Binary and thinned character

### B. Feature Extraction

Every character may be identified by its geometric specifications such as corners, endings and bifurcations. We have tried to sum these techniques. Endings may be classified as points where the character ends that is no more connected pixels can be identified beyond this point. Corners are of specific concern as they are points before and after which the gradient of the pattern changes. Bifurcations are the types of pixels where a single line gives rise to two other lines. In simple words, bifurcations may be defined as meeting point of two lines. Based on these three set of features and finding the number of pixels in each image, we use a rule based classifier which compares the features from input character with a well establish database of features to find the exact character.

*1) Corners:* Detection of corners in a character also helps in the exact detection of the character. Many alphabets such as "M", "N" and "K" have a very distinctive shape consisting of corners in a distinct pattern. Figure 3 shows different structures for corners within a 3×3 window.

*2) Endings:* Endings are the end points of lines in a character and they may be detected using simple geometric approaches. One way of ending detection may be to use a simple mask and do the calculations on the basis of this mask. Figure 4 shows different structures for endings within a 3×3 mask.

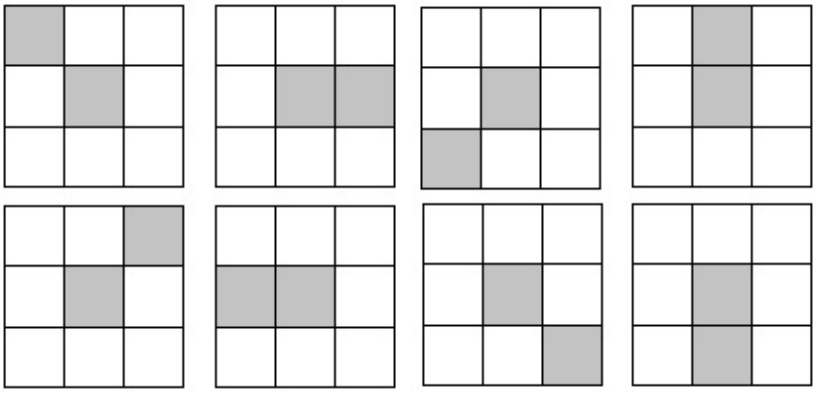

Fig. 4. Different possible structures for endings in 3×3 mask

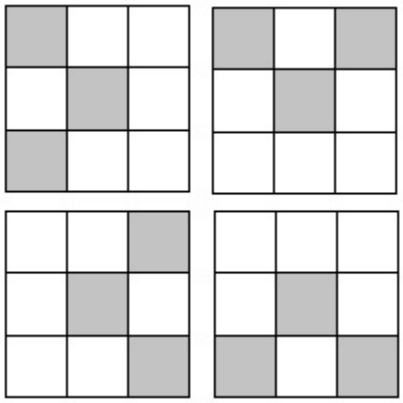

Fig. 3. Different possible structures for corners in 3×3 mask

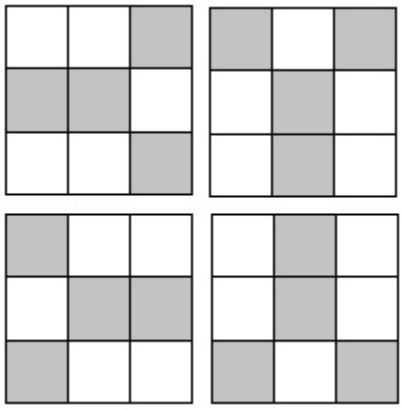

Fig. 5. Different possible structures for bifurcations in 3×3 mask

*3) Bifurcations:* Bifurcations are the pixels, beyond which a single line is divided into two or more. For the problem of optical character recognition, we only consider the bifurcation points which give rise to two other lines. Reason being that all the characters include bifurcations in which only two lines originate. Determining the type and nature of the bifurcation we may classify the character under consideration. Figure 5 shows different structures for bifurcations.

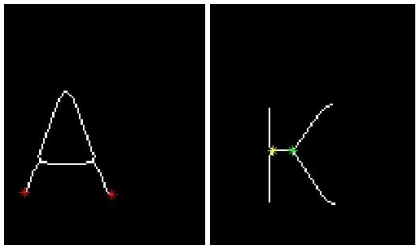

Fig. 6. Feature extraction. Endings and bifurcations for character $A$ and $K$ respectively

We have used crossing number to extract all features as given in equation and finds edges by applying crossing number method given in equation 6 [14]

$$C(p) = \frac{1}{2}\sum_{i=1}^{8} |f(p_{imod8}) - f(p_{i-1})| \tag{6}$$

Where $p_0$ to $p_7$ are the character pixels belonging to an clockwise ordered sequence of pixels defining the 8-neighborhood of p and $f(p)$ is the pixel value in thinned character. $f(p) = 1$ for character pixels and zero elsewhere. $C(p) = 1$, $C(p) = 2$ and $C(p) = 3$ correspond to character ending, corner and bifurcation respectively. Figure 6 shows the extracted endings and bifurcation for two characters.

## III. Results

The proposed system is evaluated using five samples for each character. Table 1 shows the value of each feature for all characters which are used for identification.

There are some characters which have same values for features such as character $N$ and $Z$ have 2 endings and 2 corners. Similarly there is confusion between $G$, $L$ and $V$ due to same feature values. In order to cater these problems, the proposed system further uses a line tracking and scanning system to lower the confusion between the characters having same feature values. The $C$ and $S$ have 2 endings but they are differentiated by the systems by calculating the transitions between both end points. In $C$ there is no transition between the endings but in case of $S$ there is one transition. Using

TABLE I
FEATURE VALUES FOR EACH CHARACTER

| Character | Corners | Endings | Bifurcations | Character | Corners | Endings | Bifurcations |
|---|---|---|---|---|---|---|---|
| A | 1 | 2 | 2 | B | 2 | 0 | 2 |
| C | 0 | 2 | 0 | D | 2 | 0 | 0 |
| E | 2 | 3 | 1 | F | 1 | 3 | 1 |
| G | 1 | 2 | 0 | H | 0 | 4 | 2 |
| I | 0 | 2 | 0 | J | 0 | 1 | 3 |
| K | 0 | 4 | 1 | L | 1 | 2 | 0 |
| M | 3 | 2 | 0 | N | 2 | 2 | 0 |
| O | 0 | 0 | 0 | P | 1 | 1 | 1 |
| Q | 0 | 1 | 1 | R | 1 | 2 | 2 |
| S | 0 | 2 | 0 | T | 0 | 3 | 1 |
| U | 0 | 2 | 0 | V | 1 | 2 | 0 |
| W | 3 | 2 | 0 | X | 0 | 4 | 0 |
| Y | 0 | 3 | 1 | Z | 2 | 2 | 0 |

TABLE II
PERFORMANCE EVALUATION OF PROPOSED SYSTEM

| Characters | Total Samples | Correctly Classify | Wrongly Classify | Accuracy (%) |
|---|---|---|---|---|
| A & R | 10 | 9 | 1 | 90 |
| G, L & V | 15 | 13 | 2 | 86.6 |
| N & Z | 10 | 4 | 6 | 40 |
| M & W | 10 | 5 | 5 | 50 |
| C, I, S & U | 20 | 17 | 3 | 85 |
| T & Y | 10 | 7 | 3 | 70 |
| Remaining Characters | 55 | 55 | 0 | 100 |
| Overall | 130 | 110 | 20 | 84.6 |

a supervised tracking method, the accuracy of the system is improved. Table-2 shows the evaluation results for proposed system. It shows in detail the accuracies for those characters having same feature values and for overall system.

## IV. CONCLUSION

The paper presented an automated system for identification of different characters as an application of optical character recognition. It applied preprocessing to remove the noise and converted the gray scaled image into binary by applying an adaptive threshold. It further applied morphological thinning operation to ease the feature extraction procedure. The feature extraction stage extracted the number of corners, ending and bifurcations from thinning character and using a simple rule based method it identifies the character. The proposed system is invariant to translation and rotation. The accuracy of proposed system can be improved further if we include the spectral analysis for each character such as maximas and minimas.

## REFERENCES


[1] S. Mori, C. Y. Suen, and K. Yanamoto, "Historical review of OCR research and development", Proceedings of the IEEE, vol. 80, no. 7, pp. 1029-1058, 1992.

[2] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey", IEEE Transactions on Pattern Analysis and Machine Vision, Vol. 22, No.1, pp. 63-84, 2000.

[3] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition - a survey", Pattern Recognition, Vol. 29, No.4, pp.641-662, 1996.

[4] G. Tauschek:Reading machine. U.S. Patent 2026329, December 1935.

[5] Kumar, S., Gupta, R., Khanna, N., Chaudhury, S., and Joshi, S., "Text extraction and document image segmentation using matched wavelets and mrf model", IEEE Transactions on Image Processing, Vol. 16, No.8, pp.2117-2128, 2007.

[6] Krempp, A., Geman, D., and Amit, Y., " Sequential learning of reusable parts for object detection", Technical report, Computer Science Department, Johns Hopkins University, 2002

[7] Clark, P., Mirmehdi, M. (2002). Recognising text in real scenes. International Journal on Document Analysis and Recognition, Vol. 4, pp.243-257, 2002

[8] Brown, M. S., Sun, M., Yang, R., yun, L., and Seales,W. B., "Restoring 2d content from distorted documents", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.

[9] D. Zhang and G. Lu, "A comparative study of three region shape descriptors", DICTA: Digital Image Computing Techniques and Applications, pp. 21-22, 2002.

[10] Plamondon, R. and Srihari, S. N., "On-line and offline handwriting recognition: A comprehensive survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp. 63-84, 2000.

[11] Pal, U., Sharma, N., Wakabayashi, T., and Kimura, F., "Off-line handwritten character recognition of devnagari script", In International Conference on Document Analysis and Recognition (ICDAR), pp. 496-500, 2007.

[12] Tu, Z., Chen, X., Yuille, A. L., and Zhu, S. C., "Image parsing: Unifying segmentation, detection, and recognition", International Journal of Computer Vision, Marr Prize Issue, 2012.

[13] A. Tariq, M. U. Akram, "An Automated System for Colored Retinal Image Background and Noise Segmentation", IEEE Symposium on Industrial Electronics and Applications (ISIEA 2010), pp. 405-409, 2010.

[14] M. U. Akram, S. A. Khan, "Multilayered Thresholding Based Blood Vessel Segmentation for Screening of Diabetic Retinopathy", Engineering with Computers (EWCO), Vol. 29, No. 2, pp. 165-173, 2013.