Character Recognition using Statistical Parameters

Nadeem Abbas Zaidi[1] **Noor Muhammad Shiekh**[2]
Electrical Engineering Department
University of Engineering & technology
GT Road, 5489, Lahore
Pakistan

Abstract: - Object recognition is an important task in many image processing and pattern recognition applications. Character recognition is one of the most successful applications of automatic pattern recognition techniques and essentially involves identification and classification of two dimensional (2D) signal structures. The ability to identify machine printed characters in an automated or a semi-automated manner has practical significance and different techniques are applied to get the desired recognition results. The process of character recognition involves several steps including pre-processing, feature extraction and classification. Pre- processing includes the removal of noise from the image, convert into binary image and then segment into individual character. After segmentation, each character is resized according to its aspect ratio and then boundary of the character is extracted. Feature extraction is done by applying central moment invariant technique to extract the relevant features. Classification is performed by minimum distance classifier in a 9-dimension Euclidian feature space

Key-Words: - Object recognition; OCR; Edge detection; Feature extraction; Classification

1 Introduction

Optical character recognition (OCR) has been a topic of interest since late 1940's when Jacob Rabinow started his work in this field [1]. OCR is the translation of optically scanned bitmaps of printed or written text characters into character codes. This is an efficient way to turn hard-copy materials into data files that can be edited and otherwise manipulated on the computers. Character recognition essentially deals with identification of two dimensional signal structures, which comprise of ideal alphanumeric characters embodied in noise. It is an important branch of machine perception. Mail delivery system, bank cheques, documents clearing, commercial forms processing to automatic translation of written material in different languages to provide reading aid are its major commercial applications. In addition, OCR technology can also be used in the social/educational arena. Applications of OCR with the blind and visually impaired allows the scanning of printed text and its recognition as is customary, but the addition of the capability of a synthesizer, the recognized text can then be spoken back in synthetic speech. Lastly and inevitably, OCR has its applications to the individual PC user as a means of increasing productivity by allowing the modification and reuse of existing information. Business cards, magazine articles, formal documents-all can be scanned and recognized to eliminate the need for rekeying this information manually. Before features are extracted from the scanned image certain preprocessing steps are required to be carried out on these images.

The block diagram of character recognition system is shown in figure 1. This paper is organizes as follows:

Section 2 describes the preprocessing steps performed on the images obtained from any source such as scanner or camera. An efficient edge detection method is proposed to trace the boundary of 2D character shapes in section 3. Feature extraction and classification is illustrated in section 3 & 4. In section 5, experimental results and conclusion is present.

2 Image Preprocessing

Image preprocessing seeks to modify and prepare the pixel values of digitized image to produce a form that is more suitable for subsequent operations within the generic model. In preprocessing operations, we take the acquired image array as an input and produce a modified image array as an output. Classification and recognition accuracy is heavily dependent on the preprocessing stage. Because of the complexity, digital image processing

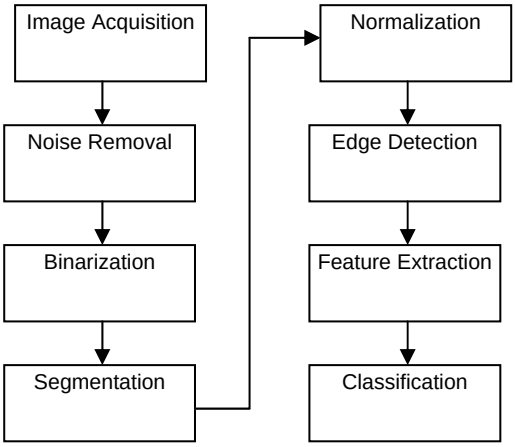

Fig. 1 Block diagram of OCR system

generally requires a large bandwidth and a number of processing steps. Processing of an image includes segmentation of an image, removing any spurious noise, binarization, image cropping, image normalization and thinning etc.

All images acquired through any acquisition system must be contaminated by a variety of noise sources. Noise is referring to stochastic variations as opposed to deterministic distortions such as shading or lack of focus. . Document image noise is due to many sources including degradation due to aging, photocopying, or during data capture, such as sensor problems in a camera, dust covering the optics or can be picked up during a point-to-point transmission of an image. Different image and signal processing methods are used to reduce noise. 'Spatial Filtering' techniques are used to reduce uncorrelated noise like 'Salt and Pepper' noise etc. In spatial filtering, an input image is passed to the function and in turn the original image has altered its appearance [2].

After noise removal from acquired document image, binerized it. Binarization is an image processing technique for converting a grayscale or color image to a binary image based upon a threshold value [3]. If a pixel in the image has an intensity value less than the threshold value, the corresponding pixel in the resultant image is to black. Otherwise, if the pixel intensity is greater than or equal to the threshold intensity, the resulting pixel is to white. Binarization is very useful for keeping the significant part of an image and getting rid of the unimportant part or noise. Since binary images are much simpler to analyze than grey scale images but the selection of the threshold T, is a critical issue. It is common to study the image histogram in order to find out threshold. In the "bimodal histogram" produced by a high contrast scene, threshold can be selected automatically by finding the lowest point of trough of the histogram.

The principle objective of the segmentation process is to partition an image into meaningful regions which corresponds to the objects of the image. This is done by systematically dividing the whole image up into constituent regions [4]. Next step is character segmentation in which document image is partitioned into individual character. This is achieved by finding the gaps between lines of the document and within the single line finding out gaps between the character, so for this purpose first lines are cropped from document and then each character is cropped from the line [5]. To detect lines of the text, take the horizontal projection of the page [6]. For this purpose, take row sum of the image. Now check this sum for non zero values. As soon as a non zero value comes, it is the starting point of the line and again as soon as the next zero comes, it is the end of the line. Hence we have find the stating and ending point of the line, so line is cropped from the image. In same the fashion all lines of the document image are cropped because there must be zero (i.e. gap) between two rows. In order to crop the characters from the line, take vertical projection. For this purpose, take column sum of the image. Now check this sum for non zero values. As soon as a non zero value comes, it is the starting point of the character and again as soon as the next zero comes, it is the end of the character. Hence we have find the stating and ending point of the character, so character is cropped from the line. In the same fashion all characters of the line are cropped because there must be zero (i.e. gap) between two characters.

After the image is segmented into characters, it is usual to adjust the size of the character region so that the following stages can assume a 'standard' character size.

3 Edge Deduction

Contours of images of objects or, in other words, edges in the paradigm of image processing and computer vision, provide valuable information towards human image understanding. The variations of image features, usually brightness, give rise to edges. More objectively, the edges are the representations of the discontinuities of image intensity function. Therefore, edge detection algorithm is essentially a process of detection of these discontinuities in an image. Edge detection is an essential tool for estimation of geometrical properties of objects in image analysis, which leads to accurate identification and recognition. Edges

form boundary or an outline of an object which segregates 2D planar shapes or objects from the background in a digital image. Boundary between overlapping objects is distinguishable, if edges in an image are identified accurately. The basic shape and topological properties of an object are present in boundary representation.

Several edge detection methods exist which include derivative operators like Robert's cross gradient operator, Prewit operator, Compass operator, Sobel operator, Canny operator and Laplacian operator etc. In this paper an efficient edge detection method is proposed to trace the boundaries of 2D character shapes. Proposed algorithm is simpler in terms of computational complexity with the rest of methods.

An approach for edge detection is proposed to detect the edges of an object or 2D character shape. The process involves detection of horizontal and vertical edges and combining them to form the complete edge detected image.

The method is based on finding out the gradient by taking the difference of adjacent pixel in binary image.

$$G_x = f(x+1, y) - f(x, y)$$
$$G_y = f(x, y+1) - f(x, y)$$
$$G \approx |G_x| + |G_y| \quad (1)$$

So when a change of pixel value from 0 to 255 or vice versa is detected, difference is stored and marked as an edge in a new 2D array. Result of proposed algorithm is the cumulative sum of horizontal and vertical edges. The proposed method only involves difference among adjacent pixel values instead of convolution mask used in other gradient methods for edge detection. This reduces the processing time due to simple computation. The other advantage of the proposed method is that the boundaries traced are thin and no additional post processing is involved as compare to other gradient methods.

4 Feature Extraction

Feature extraction is an important step in achieving best performance with respect to the implementation of a character recognizer, which will have high efficiency, and low error rate for the recognition of characters. Feature extraction means to determine various attributes as well as properties associated with a region or object. The objective of feature extraction is to represent an object in a compact way that facilitate image analysis task in terms of algorithmic simplicity and computational efficiency. Selection of feature extraction methods is probably the most important factor in achieving high recognition performance. The extracted features must be invariant to the expected distortions and variations that a character may have in a specific application. Dimensionality of feature selection is another problem which normally a feature based recognizer has to face. This puts on constrained to use limited training set and hence features must be kept reasonably small, if statistical classifier is used.

Different feature detection techniques have been employed. Initially template matching was used to find the whole character as a feature, while later; sub features of the character were sought. Algorithm found the boundary outlines, the character skeleton or medial axis, the Fourier [7] or Wavelet coefficients of the spatial pattern, various moments both spatial and grey-level, and topological properties such as the number of holes in a pattern. All have been used as features for classification of the character regions. Different feature extraction methods are designed for different representations of character, such as solid binary characters, character contour with traced edges, character skeletons or gray level sub images of each individual character. Most methods were specific to a given set of characters, i.e. only the digits, only the upper case letters of a given font, etc. For example the crossing method was developed only for use on digits. By the same time, however, more robust feature extraction schemes have been emerged such as, Zoning, Distance measurement, Fourier and Moment analysis. Moments are features based on shape characteristics or properties. Method of 'Geometric Moment Invariants' is statistical feature measurement method that describes the properties of 2D object shape.

In this paper 'Geometric Moments Invariant' technique is used to extract relevant features of a character. Moment descriptors have been developed as features in pattern recognition and describe the properties of 2D shapes related to its PDF [8].

Measurements using carefully selected moment functions can ensure that the extracted features are invariant under translation, scale change, and rotation. More importantly, moments can represent each character uniquely regardless of how close the characters are in terms of local features. This unique nature of moments makes the

method of moments an appropriate candidate for feature extraction in character recognition [9]. In the feature extraction stage, moments were computed including measurements on central moments with reference to centroid of the particular character shape. Feature vector comprising mean values of central moments, compactness and eccentricity is a 9-dimensional feature vector in a Euclidian feature space.

5 Object Classification

Together with feature extraction, the most crucial step in the process of pattern recognition is classification. Classification or recognition is a process of correctly assigning unknown patterns to their respective pattern class. Ideally this classification process should be completely automatic and error free. Classification is concerned with making decisions concerning the class membership of a pattern in question. All preceding stages should be designed and tuned for improving its success. The output of the classifier may either be a discrete selection of one of the predefined classes, or a real-valued vector expressing the values of the pattern originated from corresponding class.

In this paper 'Classification' is performed by using Minimum Distance Classifier in 9-dimensional Euclidian Feature Space. Training of classifier was accomplished by employing supervised learning and is based on the data matrix of dimensions (62 x 9) comprising mean values of central moments, compactness and eccentricity used for training. Classifier has been tested on the total of 62 characters including printed Capital, Small and Numerals of English letters in different font. Different size variations were used from font size '12' to font size '48' in classifier learning. The Class label is assigned to a particular class of characters to which a pattern belongs and having closer minimum Euclidian distance between the test and training feature vector.

6 Conclusion

Classification of pattern extracted by moment analysis is performed in '9-dimensional Euclidian' feature space utilizing 'Minimum Distance Classifier'. The developed algorithm is computationally efficient than template matching technique which is computationally heavy. For character recognition based on boundary tracing approach, overall recognition accuracy of isolated printed character images including capital, small and numerals in different font for test set was 97 %. This system has purely statistical approach while most of the OCRs use artificial intelligence like neural network, C-mean, Kohenan map etc, so their performance is purely based on training due to which they are dynamic while the results of this algorithm do not depend upon order of training.

The recognition efficiency of the algorithm can be improved by exploring other techniques such as 'Zernike Moments' and 'Wavelet Descriptors'. The applications of recognition methodologies have great impact in designing any system with optimal performance. The scope is not limited, with the advancement in the current methodologies and systems, character recognition is considered to be advantageous future aspect.

References

[1] S. Mori, C. Y. Suen, and K. Yamamot, "Historical review of OCR research and development." *Proceedings of the IEEE*, 80(7):1029--1058, July 1992.

[2] Earl Gose and Richard Johnsonbaugh, "Pattern Recognition and Image Analysis", Prentice Hall, Ed. New Delhi, 2003

[3] Qivind Due Trier and A. K. Jain, "Goal directed evaluation of binarization methods", *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1191–1201, 1995.

[4] O'Gorman L, "The document spectrum for structural page layout analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1162–1173, 1993.

[5] Yilu, "Machine printed Character Segmentation-An Overview", *Pattern Recognition*, pp. 67-80, 1994.

[6] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques", Academic Press, New York, pp. 100-132, 1985.

[7] Reti and Czinege, "Generlazed Fouriere Descriptors for Shape Analysis of 3D Closed Curved", *Acta Stereol*, Vol. 12, pp. 95-102, 1993.

[8] T. H. Reiss, "Revised fundamental theorem of moment invariants", *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 806–834, 1991.

[9] M.K.Hu, "Visual Pattern Recognition by Moment Invariants", *IRE Trans. Information Theory*, Vol. IT-8, pp. 179-187, 1962.