Neural Network Approach for Character Recognition and Text Detection: A Survey

Sachin Kumar1, H. S. Bhadauria2, Annapurna Singh3

1,2,3Department of Computer Science & Engineering, G.B. Pant Engineering College, Pauri, Uttarakhand, INDIA

ABSTRACT
Text detection and Character recognition from the image has been one of the most interesting and challenging research areas in field of pattern recognition, artificial intelligence, machine vision and image processing in the recent years. There are basically four steps that include preprocessing, feature extraction, candidate's selection, and desired character recognition to develop any of the character recognition system.

Optical character recognition is the technique to convert text from the image into computer or machine readable form. Like intelligent character recognition (ICR) one character is taken at a time to make it editable by the machine. There are several approaches for developing the OCR system but in this review we emphasis on OCR using artificial neural network trained by back propagation algorithm and fuzzy logic.

Keywords - Character recognition, OCR, Artificial neural network, Fuzzy logic.

I. INTRODUCTION

Optical Character Recognition (OCR) is the process of translating images of printed text, captured text, typewritten or handwritten text into a format understood by machines for the purpose of indexing, searching, editing and a reduction in storage size [1]. The recognition of characters from scanned images of documents has been a problem that has received much attention in the fields of artificial intelligence, pattern recognition and classification, digital image processing, etc. There are several approaches for developing the OCR system each of them comprises of two-step schema: (a) represent the character as a vector of features and (b) classify the feature vector into classes [2]. Selection of a feature extraction method is important in achieving high recognition performance. A feature extraction algorithm must be robust enough so that for a variety of instances of the same symbol, similar feature sets are generated, thereby making the subsequent classification task less difficult [3]. On the other hand, Vapnik et al. [4] have suggested that powerful classifications algorithms suffice even when given features are just sufficiently discriminative.

There have been quite a number of successes in determination of invariant features in hand writing and a wide range of classification methods have been extensively researched. However, as mentioned in, most character recognition techniques use a “one model fits all” appeal, which states that set of features and a classification method are developed and every test pattern is subjected to the same process regardless of the constraints present in the area of expertise. It is shown that approaches which employ a hierarchical treatment of patterns can have considerable advantages compared to the “one model fits all” appeal, not only improving the recognition accuracy but also reducing the computational cost as well. In Park et al.[5], a dynamic character recognizer were implemented. This recognizer begins with features extracted in a coarse resolution and focuses on smaller sub-images of the character on each recursive pass, thus working with a finer resolution of a sub-image each span of time till classification meets acceptance criterion. By employing an approach called gaze planning which is defined as an expansion of only some of the nodes in a tree structure similar to quad trees [6], not all of the sub-images are subjected to further sub division but only those where it is believed that features of interest are present. So, a feature vector is extracted for each character that has more information from those sub-images that are deemed to be more important than others. The feature vector is generated by combing all features extracted in each sub-image. These features are based on histogram of gradient and moment-based projections. In [7] the character image is sub divided recursively into smaller sub-images based on the quad tree rule. The input image is then represented by fractal codes obtained at each iteration by encoding algorithm. In [8] a feature extraction technique relied on recursive sub divisions of the image for the recognition of mathematical glyphs is introduced. Each split is based on the Centre of gravity of the corresponding sub-image.

The initial splitting is vertical and each level of splitting then alternates between horizontal and vertical. For each rectangular region a four dimensional feature vector is extracted consisting of the vertical or horizontal component of the centroid and the three second order central moments. Moreover, other approaches focus on measuring the similarity dissimilarity between shapes by mapping one character.
on to another. In Belongie et al. [9] the shape context is presented. Each shape is represented by a set of points extracted from the contour or outline. For each shape, a descriptor is introduced, the shape context, which is the log-polar histogram of the point. Corresponding points on two similar shapes are supposed to have the same shape context thus resulting in a bipartite graph matching problem. In [10] two characters are matched by deforming the outline of one to fit the edge strengths of the other, and an error in similarity is originate from the amount of deformation needed, the goodness off it of the edges and the interior overlap between the deformed shapes.

Most classification strategies in OCR deal with a large number of classes striving to detect the best hyperplane among them. However, such approaches are vulnerable to classification errors when classes of similar shapes are present since they are not easily distinguished. In [11] a two-stage classification approach is presented to detect and solve possible conflicts between characters such as ‘N’ and ‘Z’ or ‘U’ and ‘V’. During the first stage, a single classifier or ensemble of classifiers detect potential conflicts. The second processing stage becomes active only when a decision on the difficult cases must be taken. A comparative study between three different two-stage hierarchical learning architectures can be found in [12].

II. OCR PREPROCESSING

Data preprocessing [13] refers how the image can be preprocessed in order to attain the good quality image and how the image can be preprocessed so as to improve the efficiency and ease of the process of data mining. There are a number of data preprocessing techniques. Data cleaning can be applied to correct inconsistencies and remove noise from the data. Data integration is used to merge data from multiple sources into a coherent data store, such as a data warehouse and repository. Data transformations, such as normalization, may be applied which may improve the efficiency and accuracy of mining algorithms involving distance measurements. Data reduction can reduce the size of data by eliminating redundant features, clustering and aggregating.

2.1 Binarization

Binarization techniques, which have been proposed especially for document binarization [14], can alternatively be used for the binarization of the identified text blocks. These techniques are preferable to be applied in cases where the text blocks contain noise that must be removed. A recent description of the five document binarization techniques included in our system is given as follows:

2.1.1 Bernsen’s technique (1986) uses a local threshold which is calculated as follows:

\[
T(x, y) = \begin{cases} 
\frac{P_{low} + P_{high}}{2}, & \text{if } P_{high} - P_{low} \geq L \\
GT, & \text{if } P_{high} - P_{low} < L
\end{cases}
\]  

(1)

where \( P_{low} \) and \( P_{high} \) refers the lowest and the highest gray level value in a \( N \times N \) window centered in the pixel \( (x, y) \), respectively and \( GT \) a global threshold value (for example a threshold value that is calculated from the application of the method of Otsu to the entire image).

2.1.2 Niblack’s technique (1986) which also uses a local threshold calculated as follows:

\[
T(x, y) = m(x, y) + ks(x, y)
\]  

(2)

where \( m(x,y) \) and \( s(x,y) \) refers to the local mean and standard deviation values in a \( N \times N \) window centered on the pixel \( (x,y) \), respectively.

2.1.3 Sauvola and Pietikainen’s technique (2000) proposes the following calculation for the local threshold:

\[
T(x, y) = m(x, y) \left[ 1 + k(1 - \frac{s(x, y)}{R}) \right]
\]  

(3)

where \( m(x,y) \) and \( s(x,y) \) refers to the same as in the previous technique and \( R \) is a constant equal to 128 in most cases.
2.2 Morphological Operators

Morphological operator [15] removes isolated specks and holes in characters. Common morphological operations are:

2.2.1 Erosion
Equation of an erosion of an image I by the structure element H is given by the set operation.

\[
I \ominus H = \{ p \in Z^2 | (p + q) \in I, \text{for every } q \in H \} \tag{4}
\]

2.2.2 Dilation
Equation of a dilation of an image I by the structure element H is given by the set operation.

\[
I \oplus H = \{ p + q | p \in I, q \in H \} \tag{5}
\]

2.2.3 Closing
Operation defined as dilation followed by erosion:

\[
I \circ H = (I \oplus H) \ominus H \tag{6}
\]

Holes in the foreground that is smaller than H will be filled.

2.2.4 Opening
Operation defined as erosion followed by dilation:

\[
I \bullet H = (I \ominus H) \oplus H \tag{7}
\]

Stray foreground structures that are smaller than the H structure element will disappear. Larger structures will remain.

2.2.5 Finding the outline of the foreground
The outline image B(u,v) of a binary object can be computed using a dilation followed by a subtraction (or XOR operation):

\[
I' = I \ominus H \tag{8}
\]

\[
B(u, v) = \text{XOR}(I'(u, v), I(u, v)) \tag{9}
\]

2.2.6 Finding the skeleton of the foreground
Often run erosion operation, stop when 1-pixel thick.

2.3 Segmentation

The goal of image segmentation[16] is to cluster pixels into salient image regions, i.e., regions belongs to individual surfaces, objects and natural parts of objects. Segmentation could be used for occlusion boundary estimation, object recognition within motion or stereo systems, image editing, image compression, or image database look-up. Here we consider bottom-up image segmentation. That is, we ignore (top-down) contributions from object recognition in the segmentation process. For input we primarily consider image brightness here, although similar techniques can be used with color, motion, and/or stereo disparity information.

III. OCR FEATURE EXTRACTION

3.1 Moment based feature

Moments describe a shape’s layout (the arrangement of its pixels), a bit like combining area, compactness, irregularity and higher order descriptions together. Moments are a global description of a shape, accruing this same advantage as Fourier descriptors since there is selectivity, which is an in-built ability to discern, and filter, noise. Further, in image analysis, they are statistical moments[17], as opposed to mechanical ones, but the two are analogous. The two-dimensional Cartesian moments associated with an order that starts from low up to higher orders. The moment of order p and q, \(m_{pq}\) of a function I(x,y) is as follow:

\[
m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q I(x,y) dx dy \tag{10}
\]

3.2 Hough and Chain code transform

The generalized Hough transform is used when the shape of the feature that we wish to isolate does not have a simple analytic equation describing its boundary or contour. In such case, instead of using a parametric equation of the curve, we prefer to use a look-up table to define the relationship between the boundary positions and orientations and the Hough parameters. (The values in the Look up table must be computed during a preliminary phase using a prototype shape.)

For example, suppose that we know the shape and orientation of the desired feature. We can specify an arbitrary reference point (\(x_{ref},y_{ref}\)) within the feature, with respect to which the shape (i.e. the distance r and angle of normal lines drawn from the boundary to this reference point \(\omega\)) of the feature is defined. Our look-up table (i.e. R-table) will consist of these distance and direction pairs, indexed by the orientation \(\omega\) of the boundary.

![Fig.2. Description of R-table components](image)

The Hough transform[18] space is now defined in terms of the possible positions of the shape in the image, i.e the possible ranges of \((x_{ref},y_{ref})\). In other words, the transformation is defined by:
\[ x_{\text{ref}} = x(\beta) r \cos(\beta) \]  
(11)

\[ y_{\text{ref}} = y(\beta) r \sin(\beta) \]  
(12)

(The \( r \) and \( \beta \) values are derived from the R-table for particular known orientations \( \omega \)). In case the orientation of the desired feature is not known, this procedure is complicated by the fact that we must extend the accumulator by incorporating an extra parameter to account for changes in orientation.

To obtain a representation of a contour, we can simply store the coordinates of a sequence of pixels in the image. Alternatively, we can just store the relative position between consecutive pixels. This is the basic idea behind chain codes. Chain codes are one of the oldest techniques in computer vision, originally introduced in the 1960s (Freeman, 1961; an excellent review came later: Freeman, 1974). Essentially, the set of pixels in the border of a shape is translated into a set of connections between them. Given a complete border, one that is a set of connected points, then starting from one pixel we need to be able to determine the direction in which the next pixel is to be found. Namely, the next pixel is one of the adjacent points in one of the major compass directions. Thus, the chain code[19] is formed by concatenating the number that designates the direction of the next pixel. That is, given a pixel, the successive direction from one pixel to the next pixel becomes an element in the final code. This is repeated for each point until the start point is reached when the (closed) shape is completely analyzed.

![Fig.3. Connectivity in chain codes](image)

### 3.3 Fourier transform and series

Fourier descriptors[20], often attributed to early work by Cosgriff (1960), allow us to bring the power of Fourier theory to shape description. The main idea is to characterize a contour by a set of numbers that represent the frequency content of a whole shape. Based on frequency analysis, we can select a small set of numbers (the Fourier coefficients) that describe a shape rather than any noise (i.e. the noise affecting the spatial position of the boundary pixels). The general recipe to obtain a Fourier description of the curve involves two main steps. First, we have to define a representation of a curve and then we expand it using Fourier theory. From which we can obtain alternative flavors by combining different curve representations and different Fourier expansions. In this review we consider Fourier descriptors of angular and complex contour representations. Nevertheless, Fourier expansions can be developed for other curve representations.

In the most basic form, the coordinates of boundary pixels are \( x \) and \( y \) point coordinates. A Fourier description gives the set of spatial frequencies that fit the points on boundary. The first element of the Fourier components (the d.c. component) is simply the average value of the \( x \) and \( y \) coordinates, giving the coordinates of the center point of the boundary. The second component essentially gives the radius of the circle that best fits the points. Accordingly, a circle can be described by its zero-order and first order components (the d.c. component and first harmonic). The higher order components increasingly describe detail, as they are associated with higher frequencies.
IV. OCR CLASSIFICATION

The classification and recognition of individual characteristics and behaviors constitute a preliminary step and is an important objective in the behavioral sciences. Here we defined a methodology based on one of the principles of artificial neural networks: the backpropagation gradient[21].

Artificial neural networks [22] were initially developed according to the elementary principle of the operation of the (human) neural system. Since then, a very large variety of networks have been constructed. All are composed of units (neurons), and connections between them, which together determine the behavior of the network. The choice of the network type depends on the problem to be solved; the backpropagation gradient network is the most frequently used (Rumelhart et al., 1986; Kosko, 1992). This network consists of three or

Fig. 4. An ellipse and its Fourier description

Fig. 5. Structure of neural network
more neuron layers: a input layer, output layer and at least single hidden layer. In most cases, a network with only one hidden layer is used to restrict calculation time, especially when the results obtained are satisfactory. All the neurons of each layer (except the neurons of the last one) are connected by an axon to each neuron of the next layer (Fig. 5).

V. CONCLUSION

Since there are various techniques and methodology to develop a system that automatically recognizes the text from the captured and drawn images. Here in this review we focused on artificial neural network trained by back propagation algorithm to develop a system called Optical character recognition system. This paper described all of the steps to develop such system with artificial neural network as a classifier by gone through various recent and previous research work on such type of systems.

REFERENCES