OPTICAL CHARACTER RECOGNITION USING NEURAL NETWORK

Krunal H. kubavat [1], Sima K. Gonsai [2]

[1,2]*Department of Electronics and Communications, L. D. College of Engineering, Gujarat Technological University Ahmedabad*
kunal.kubavat23@gmail.com
simagonsai@gmail.com

Abstract--Recognizing characters with machines has started gaining importance over the years especially handwritten characters. Advancements in image processing methods lead to generation of better feature extraction methods. These features are then applied to some classifier algorithms like K-NN classifier, neural networks to identify characters. This paper examines the statistical method available for character recognition. Some applications of character recognition are aid for blind, automatic number-plate readers, signature verification and identification, mail sorting, Data entry.

Keywords : K-NN classifier, Neural network.

I. INTRODUCTION

OCR belongs to the family of techniques that performing automatic identification. Optical Character Recognition handles the problem that recognizes optically processed characters. Optical recognition can be performed off-line after the writing or printing has been completed, while the on-line recognition where the computer recognizes the characters as they are drawn. Performance of recognition for both hand written and printed characters depends on the accuracy of input data.

Better performance can be reached with accurate input. On the other side, when it comes to totally unconstrained handwritten data, OCR machines are still a long way from reading as good as humans. However, the computer reads faster than human and technical advances are gradually trying to bring the technology closer to its ideal.

II. METHODS OF OCR

In automatic recognition of patterns, the main principle is first to make the machine learn which classes of patterns that may occur and what they may look like. In OCR the patterns may be letters, numbers and some special symbols like commas, question marks etc., while the different classes represent the different characters. The learning process of the machine is done by showing the machine examples of characters related to all the different classes. Based upon these examples the machine builds a description or a prototype of each class of characters. The unknown characters are compared with the previously obtained descriptions, during the process of recognition, and assigned the class that provides the best match. Some model does provide facilities for training in the case of new type of characters are available..

III. COMPONENTS OF AN OCR SYSTEM

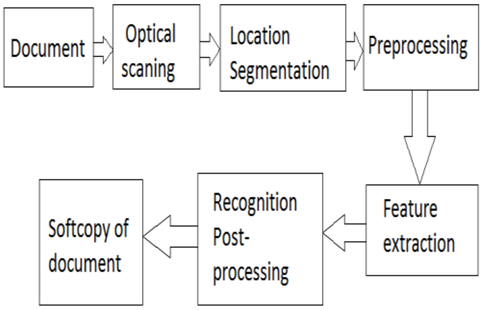

Figure 1: Components of an OCR system

A typical OCR system consists of several processes. In figure 1 a common setup is illustrated. The first block in the process is aimed to digitize the analog data using an optical scanner. Segmentation process helps to extract each symbol, where area containing text is located. The extracted symbols may then be preprocessed to facilitate the extraction of features in the next process.

The identity of each symbol is obtained by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning process. At last contextual information is used to rebuild the words and numbers of the original document. These steps and some of the methods involved are explained in more detail in the next sections.

A .Optical scanning.

By scanning process a digital image of the original data is captured. Optical scanners consist of a transport mechanism and a sensing device that converts light intensity into gray-levels are used here. Printed documents are usually consisting of black and white format. That is why, while performing OCR it is natural to convert the multilevel image into a bi-level image of black and white. This process, known as thresholding, helps to reduce size and computational efforts.

B. Location and segmentation

Segmentation is a process that finds the constituents of an image. The region of document must be found where data have been printed and distinguish them from figures and graphics. For example, when doing automatic mail sorting, the address must be located and separated from other data on the envelope like stamps and company logos, before recognition.

When applied to text, segmentation is the differentiation of characters or words. Words are isolated into character in almost all the character recognition algorithm. Generally this segmentation is performed by isolating each connected component. This method is can be implemented easily, but if characters are touched or fragmented and consist of several parts, problem may occur.

C. Preprocessing

Certain amount of noise may be there in the image given by optical scanning. The resolution on the scanner and the success of the applied technique for thresholding, determine the need of the characters to be smeared or broken. Some defects, which may lead to poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters. The smoothing consists of both filling and thinning. Small breaks, gaps and holes in the digitized characters are removed through filling, while width of the line is reduced through thinning. In the most common technique for smoothing a window moves across the binary image of the character, applying certain rules to the contents of the window. After smoothing, preprocessing usually use normalization. Characters of uniform size, slant and rotation are extracted using normalization. After finding the angle of rotation we can correct rotation. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew. We can find the rotation angle of a single symbol only after the symbol has been recognized.

D.Feature extraction

The purpose of feature extraction is to obtain the essential characteristics of the symbols, and it is concluded that this is one of the most difficult problem of pattern recognition. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into five main groups.

1) Template Matching and Correlation Technique

In 1929 Tausheck obtained a patent on OCR in Germany and this was the first conceived idea of an OCR. Their approach was what is referred to as template matching in the literature. The template matching process can be roughly divided into two sub processes, i.e. superimposing an input shape on a template and measuring the degree of coincidence between the input shape and the template. The template, which matches most closely with the unknown, provides recognition. The two-dimensional template matching is very sensitive to noise and difficult to adapt to a different font. A variation of template matching approach is to test only selected pixels and employ a decision tree for further analysis. Peephole method is one of the simplest methods based on selected pixels matching approach. In this approach, the main difficulty lies in selecting the invariant discriminating set of pixels for the alphabet. Moreover, from an Artificial Intelligence perspective, template matching has been ruled out as an explanation for human performance.

2) Features Derived from the Statistical Distribution of Points

This technique is based on matching on feature planes or spaces, which are distributed on an n-dimensional plane where n is the number of features. This approach is referred to as statistical or decision theoretic approach. Unlike template matching where an input character is directly compared with a standard set of stored prototypes. Many samples of a pattern are used for collecting statistics. This phase is known as the training phase. The objective is to expose the system to natural variants of a character. Recognition process uses this statistics for identifying an unknown character. The objective is to expose the system to natural variants of a character. The recognition process uses this statistics for partitioning the feature space. For instance, in the K-L expansion one of the first attempt in statistical feature extraction, orthogonal vectors are generated from a data set. For the vectors, the covariance matrix is constructed and its eigenvectors are solved which form the coordinates of the given pattern space. Initially, the correlation was pixel-based which led to large number of covariance matrices. This approach was further refined to the use of class-based correlation instead of pixel-based one which led to compact space size. However, this approach was very sensitive to noise and variation in stroke thickness. To make the approach tolerant to variation and noise, a tree structure was used for making a decision and multiple prototypes were stored for each class. Researchers for classification have used the Fourier series expansions, Walsh, Haar, and Hadamard series expansion.

$$\text{Mean}\ \mu = \sum x \sum P(x, y) \tag{1}$$

$$\text{Variance} = \sum (x-\mu) 2\, P(x, y) \tag{2}$$

These parameters are used in calculations.

3) Geometrical and Topological Features

The classifier is expected to recognize the natural variants of a character but discriminate between similar looking characters such as ‘k’ – ‘ph’, ‘p’ - ‘Sh’ etc. This is a contradicting requirement which makes the classification task challenging. The structural approach has the capability of meeting this requirement. The multiple prototypes are stored for each class, to take care of the natural variants of the character. However, a large number of prototypes for the same class are required to cover the natural variants when the prototypes are generated automatically. In contrast, the descriptions may be handcrafted and a suitable matching strategy incorporating expected variations is relied upon to yield the true class. The matching strategies include dynamic programming, test for isomorphism, inexact matching, relaxation techniques and multiple to-one matching. Rocha have used a conceptual model of variations and noise along with multiple to one mapping. Yet another class of structural approach is to use a phrase structured grammar for prototype descriptions and parse the unknown pattern syntactically using the grammar. Here the terminal symbols of the grammar are the primitives of strokes and non-terminals represent the pattern-classes. The production rules give the spatial relationships of the constituent primitives.

4) Hybrid Approach

Statistical approach and structural approach both have their advantages and shortcomings. The statistical features are more tolerant to noise (provided the sample space over which training has been performed is representative and realistic) than structural descriptions. The variation due to font or writing style can be more easily abstracted in structural descriptions. Two approaches are complimentary in terms of their strengths and have been combined. The primitives have to be ultimately classified using a statistical approach. Combine the approaches by mapping variable length, unordered sets of geometrical shapes to fixed length numerical vectors. This approach, the hybrid approach, has been used for Omni font, variable size character recognition systems.

5) Neural Networks

In the beginning, character recognition was regarded as a problem, which could be easily solved. But the problem turned out to be more challenging than the expectations of most of the researchers in this field. The challenge still exists and an unconstrained document recognition system matching human performance is still nowhere in the sight. The performance of a system deteriorates very rapidly with deterioration in the quality of the input or with the introduction of new font handwriting. Training phase aims at exposing the system to a large number of fonts and their natural variants. The neural networks are based on the theory of learning from the known inputs. A back propagation neural network is composed of several layers of interconnected elements. Each element computes an output, which is a function of weighted sum of its inputs. The weights are modified until a desired output is obtained. The neural networks have been employed for character recognition with varying degree of success. The neural networks are employed for integrating the results of the classifiers by adjusting weights to obtain desired output. The main weakness of the systems based on neural networks is their poor capability for generality. There is always a chance of under training or over training the system. Besides this, a neural network does not provide structural description, which is vital from artificial intelligence viewpoint. The neural network approach has solved the problem of character classification no more than the earlier described approaches. The recent research results call for the use of multiple features and intelligent ways of combining them. The combination of potentially conflicting decisions by multiple classifiers should take advantage of the strength of the individual classifier, avoid their weaknesses and improve the classification accuracy. The intersection and union of decision regions are the two most obvious methods for classification combination.

Artificial Neural Networks

One type of network sees the nodes as 'artificial neurons'. These are called artificial neural networks (ANNs). Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons.

Multilayer feed forward network

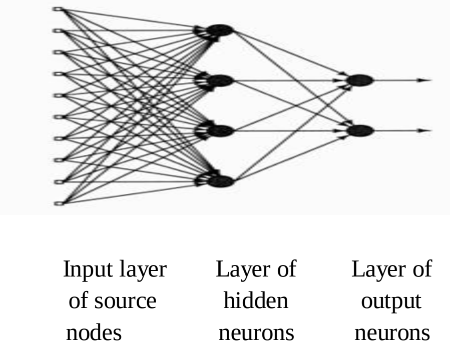

Figure 2: A 10-4-2 fully connected feed forward network

The source nodes supply respective elements of the activation pattern in the input layer of the network, which contains the input signals which should be applied to the neurons in the second layer. The output signals from second layer are used as inputs to the next (or third) layer, and go on until the last layer of the network. The neurons in each layer of the network have their inputs are the output signals of the previous layer only. The output signals set of the neurons in the output layer of the network contains the overall response of the network to the activation pattern given by the source nodes in the input layer. The network shown above Figure 2 is referred to as a 10-4-2 network in that it has 10 source nodes, 4 hidden neurons, and 2 output neurons. In another example, a feed forward network with p no. of source nodes, h1 no. of neurons in the first hidden layer, h2 no. of neurons in the second layer, and q no. of neurons in the output layer is identified as a p-hl-h2-q network. A fully connected network is one in which all the nodes in each layer of the network is connected to all the other nodes in the adjacent forward layer. In partially connected network some of the communication links (synaptic connections) are absent from the network.

Table 1:- Comparison between Template Matching and Neural Network

Type of Data	Parameter used	Accuracy (%)
Printed	Mean	100
	Variance	100
	ASM	100
	Mean,Variance,ASM	100
Handwritten	Mean	72
	Variance	78
	ASM	75
	Mean,Variance,ASM	81

IV. CONCLUSION

Optical character recognition methods are aimed to recognize documents. Various methods discussed shows different recognition rate and error rate. Template matching is most efficient way for printed characters. Use of geometrical and topological features proved to be good in some cases. In case of number plate recognition, Statistical distribution of points gives better result. In handwritten character recognition, neural networks classifier combined with statistical method gives good result. Yet most efficient method can't be pickup by just comparing recognition rates. Because varieties in handwritten characters are so diverse, so hybrid approach combining two or more methods may be suitable solution for OCR.

REFERRENCES

[1] Omar Noori, Sharifah Mumtazah Syed Ahmed and Asma Shakil, 2011 "Offline Malay Handwritten Cheque Words Recognition using Artificial Neural Network" Journal of Applied Science 11(1): 86-95.

[2] Line Eikvil, 2003. PhD Thesis on "Optical Character recognition", Norsk Regnesentral, Oslo.

[3] Seethalakshmi R., Sreeranjani T.R., Balachandar T,Sept 2005."Optical Character Recognition for printed Tamil text using Unicode" Journal of Zhejiang University Science, pp. 1297-1305.

[4] LTG (Language Technologies Group), 2003. Optical Character Recognition for Printed Kannada Text Documents SERC, IISc Bangalore.

[5] VijayaKumar B., 2001. Machine Recognition of PrintedKannada Text, IISc Bangalore,The Unicode Standard Version 3.0, Addison Wesley.

[6] Gonzalez, R.C., Woods, R.E., Eddins, S.L., 2004. Digital Image Processing Using MATLAB PHI Pearson

[8] Carbonnel, S. and E. Anquetil, 2003. Lexical post processing optimization for handwritten word recognition,7th International Conference Document Analysis Recognition, 1:477-481.

[9] Gorgel, P. and O. Oztas, 2007. Handwritten character recognition system using artificial neural networks. J. Elect. Electronics Eng., 7: 309-313.

[10] Carbonnel, S. and E. Anquetil, 2003. Lexical post-processing optimization for handwritten word recognition. Proc. 7th Int. Conf. Document Analysis Recognition, 1 : 477-481.

[11] Plamondon, Rand S.N. Srihari, 2000. On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. Pattern Anal. Mach., 22: 63-84.