A Study of Optical Character Patterns identified by the different OCR Algorithms

Purna Vithlani[*], Dr. C. K. Kumbharana[**]

[*] Research Scholar, Department of computer science, Saurashtra University, Rajkot, India.
[**] Head, Department of computer science, Saurashtra University, Rajkot, India.

Abstract- Optical Character Recognition (OCR) is a technology that provides a full alphanumeric recognition of printed or handwritten characters. Optical Character Recognition is one of the most interesting and challenging research areas in the field of Image processing. Image Acquisition, Pre-processing, Segmentation, Feature Extraction and Classification are stages of OCR. In this paper, how character patterns are identified in the classification stage by different algorithms is presented. Template Matching Algorithm, statistical Algorithm, Structural Algorithm, Neural Network Algorithm and Support Vector Machine Algorithm are presented in this paper.

Index Terms- Neural Network Algorithm, Optical Character Recognition, Statistical Algorithm, Structural Algorithm, Support Vector Machine, Template Matching.

I. INTRODUCTION

Optical Character Recognition (OCR) is a process of converting scanned document into text document so it becomes editable and searchable. OCR is the mechanical or electronic translation of images of handwritten or printed text into machine-editable text. Recognition engine of the OCR system interpret the scanned images and turn images of handwritten or printed characters into ASCII data (Machine readable characters).

An image is passes from number of stages like Image Acquisition, pre-processing, segmentation, feature extraction and classification to perform OCR. Images of OCR system might be acquired by scanning document or by capturing photograph of document in Image Acquisition stage. Pre-processing is necessary to modify the raw data to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor. Pre-processing step involves binarization, noise removal, normalization etc. Segmentation is the process of separating lines, words and characters from image. Feature extraction can be considered as finding a set of features that define the shape of the underlying character as precisely and uniquely as possible. Being the most important step of the recognition process, selection of the features is the important factor in achieving the high recognition performance. Classification stage uses the features extracted in the feature extraction stage to identify the character. The classification stage is the decision making part of a recognition system. Classification is the part of the OCR which finally recognizes individual characters and outputs them in machine editable form.

II. PREVIOUS WORK

Faisal Mohammad et.al. [1] have presented pattern matching algorithm for typewritten and handwritten characters. The binary image is divided into 5 tracks and 8 sectors. The track-sector matrix is then matched with existing template. The existing template consists of each track-sector intersection value, each track value and each sector value. If all these parameters are found to match with the template values then the resultant is the character identified.

Mr. Danish Nadeem and Miss. Saleha Rizvi [2] have proposed typewritten/handwritten character recognition using template matching. The aim is to produce a system that classifies a given input as belonging to a certain class rather than to identify them uniquely, as every input pattern. The system performs character recognition by quantification of the character into a mathematical vector entity using the geometrical properties of the character image. Recognition rate of typewritten Standard English alphabets fonts is 94.30%, typewritten Unknown English alphabets fonts is 88.02% and handwritten English alphabets is 75.42%.

Rajib et.al. [3] have proposed Hidden Markov Model based system for English Handwritten character recognition. They have employed global as well as local feature extraction methods. HMM is trained using these feature and experiment is carried out. They have created a data-base of 13000 samples collected from 100 writers written five times for each character. 2600 samples have been used to train HMM and the rest are used to test recognition model. The recognition rate is achieved 98.26% using proposed system.

Pritpal Singh and Sumit Budhiraja [4] have proposed K Nearest Neighbour algorithm to recognise handwritten Gurumukhi script. They calculate the Euclidean distance between the test point and all the reference points in order to find K nearest neighbours, and then arrange the distances in ascending order and take the reference points corresponding to the k smallest Euclidean distances. A test sample is then attributed the same class label as the label of the majority of its K nearest (reference) neighbours. The recognition rate is achieved by proposed system is 72.54%.

Jonathan J. Hull, Alan Commike and Tin-Kam HO [5] have presented Structural analysis algorithm. The structural

analysis algorithm has been fully implemented and tested in the proposed work.

Parveen Kumar, Nitin Sharma and Arun Rana [6] have proposed handwritten character recognition system using multi layer feed forward back propagation neural network without feature extraction. For the neural network, each character is resized into 70x50 pixels, which is directly subjected to training. That is, each resized character has 3500 pixels and these pixels are taken as features for training the neural network. The proposed system has 4 layers – one input layer, one output layer and two hidden layers, having 200 neurons in the first hidden layer and 100 neurons in the second hidden layer. The recognition rate is 80.96%.

Yusuf Perwej and Ashish Chaturvedi [7] have proposed neural networks for developing a system that can recognize handwritten English alphabets. Each English alphabet is represented by binary values that are used as input to a simple feature extraction system, whose output is fed to neural network system. The recognition rate is 82.5%.

Akash Ali et al. [8] have proposed handwritten Bangla character recognition using Back propagation Feed-forward neural network. First, Create binary image then, extract the feature and form input vector. Then, apply the input vector in the neural network. The experimental result shows that the proposed recognition method gives 84% accuracy and less computational cost than other method.

Nasien et al. [9] have proposed a recognition model for English handwritten (lowercase, uppercase and letter) character recognition that uses Freeman chain code (FCC) as the representation technique of an image character. Support vector machine (SVM) has been chosen for the classification. The proposed recognition model, built from SVM classifiers was efficient enough to show that applying the proposed model, a relatively higher accuracy of 98.7% for the problem of English handwritten recognition was reached [10].

Anshuman Sharma [11] has proposed handwritten digit Recognition using Support Vector Machine. In this proposed work the SVM (binary classifier) is applied to multi class numeral recognition problem by using one-versus-rest type method. The SVM is trained with the training samples using linear kernel.

III. OCR ALGORITHMS

In Feature Extraction stage, unique features of the characters are identified. These features used in classification stage to identify the character. So, Classification is called decision making stage of the OCR process. Here, how optical character patterns identified by different OCR Algorithms in classification stage is presented.

A. Template Matching Algorithm

Template matching algorithm is also known as pattern matching algorithm. Template matching is a system prototype that useful to recognize the character or alphabet by comparing two images. Template matching is the process of finding the location of sub image called a template inside an image. Once a number of corresponding templates is found their centers are used as corresponding points to determine the registration parameters. Template matching involves determining similarities between a given template and windows of the same size in an image and identifying the window that produces the highest similarity measure. In Template matching, the character itself is used as a “feature vector”.

Template matching algorithm simply identify the character by comparing character patterns with already stored template.

B. Statistical Algorithm

The purpose of the statistical algorithms is to determine to which category the given pattern belongs. By making observations and measurement processes, a set of numbers is prepared, which is used to prepare a measurement vector [12]. Statistical algorithm uses the statistical decision functions and a set of optimality criteria which maximizes the probability of the observed pattern given the model of a certain class.

Statistical algorihtms are mostly based on three major assumptions:

1. Distribution of the feature set.
2. There are sufficient statistics available for each class.
3. Collection of images to extract a set of features which represents each distinct class of patterns.

Statistical Methods

1. K-Nearest Neighbour

The k-Nearest Neighbors algorithm (k-NN) is a non-parametric method used for classification. The input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to class of that single nearest neighbor [13]. The idea behind k-Nearest Neighbor algorithm is quite straightforward. To classify a new character, the system finds the k nearest neighbors among the training datasets, and uses the categories of the k nearest neighbors to weight the category candidates [14].

2. Clustering Analysis

The goal of a clustering analysis is to divide a given set of data or objects into a cluster, which represents subsets or a group. The partition should have two properties. Homogeneity inside clusters: the data, which belongs to one cluster, should be as similar as possible. Heterogeneity between the clusters: the data, which belongs to different clusters, should be as different as possible [15]. Thus, the characters with similar features are in one cluster. Thus, in recognition process, the cluster is identified first and then the actual character.

3. Hidden Markov Modeling

A hidden markov model(HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. The Hidden Markov Model is a finite set of states,each of which is associatd with a probability

distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are "hidden" to the outside; hence the name Hidden Markov Model [16].

Mathematically Hidden Markov Model contains five elements.
1. Internal States: These states are hidden and give the flexibility to model different applications. Although they are hidden, usually there is some kind of relation between the physical significance to hidden states.
2. Output: O = {O1, O2, O3, . . . , On} an output observation alphabet.
3. Transition Probability Distribution: A = a_{ij} is a matrix. The matrix defines what the probability to transition from one state to another is.
4. Output Observation: Probability Distribution B = b_i (k) is probability of generating observation symbol o (k) while entering to state i is entered.
5. The initial state distribution (π = { πi }) is the distribution of states before jumping into any state.

Here all three symbols represents probability distributions i.e. A, B and π. The probability distributions A, B and π are usually written in HMM as a compact form denoted by lambda as λ = (A, B, π) [17].

C. Structural Algorithm

The recursive description of a complex pattern in terms of simpler patterns based on the shape of the object was the initial idea behind the creation of structural pattern recognition [18].

Structural algorihtm classifies the input patterns on the basis of components of the characters and the relationship among these components. Firstly the primitives of the character are identified and then strings of the primitives are checked on the basis of pre-decided rules [12]. Structural pattern recognition is intuitively appealing because in addition to classification, this approach also provides a description of how the given path constructed from the primitives [19]. Generally a character is represented as a production rules structure, whose left-hand side represents character labels and whose right-hand side represents string of primitives. The right-hand side of rules is compared to the string of primitives extracted from a word. So classifying a character means finding a path to a leaf [12].

D. Neural Network Algorithm

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems [20]. A neural network is a powerful data modeling tool that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain.

Neural Network consists three layers - input layer, hidden layer (optional) and output layer.

1. Input Layer

The input layer is the conduit through which the external environment presents a pattern to the neural network. Input layer take input from the external world and encode it into a convenient form. Every input neuron should represent some independent variable that has an influence over the output of the neural network.

The number of input neurons for the OCR is the number of pixels that might represent any given character. A character which represents by a 5*7 grids has 35 pixels. So it has 35 input neurons.

2. Hidden Layer

Hidden layer can't see or act upon the outside world directly. These inter-neurons communicate only with other neurons. Deciding the number of neurons in the hidden layers is a very important part of deciding your overall neural network architecture. Both the number of hidden layers and the number of neurons in each of these hidden layers must be carefully considered [21].

3. Output Layer

The output layer of the neural network is what actually presents a pattern to the external environment. The number of output neurons should be directly related to the type of work that the neural network is to perform.

The number of output neurons used by the OCR program will vary depending on how many characters the program has been trained to recognize. The default training file that is provided with the OCR program is used to train it to recognize 26 characters. Using this file, the neural network will have 26 output neurons.

Neural network algorithm identifies the character by training the neural network. Feed forward Neural Network, Feedback neural network and Self Organizing Map are the types of neural network.

E. Support Vector Machine

Support vector machines (SVMs also support vector networks) are a set of related supervised learning methods used for classification.

SVMs are relatively new approach compared to other supervised classification algorithms, they are based on statistical learning theory developed by the Russian scientist Vladimir Naumovich Vapnik back in 1962 and since then, his original ideas have been perfected by a series of new techniques and algorithms [22].

Support vector machines have proved to achieve good generalization performance with no prior knowledge of the data. The principle of an SVM is to map the input data onto a higher dimensional feature space nonlinearly related to the input space

and determine a separating hyperplane with maximum margin between the two classes in the feature space[23] [24]. This approach, in general, guarantees that the larger the margin is the lower is the generalization error of the classifier [25].

SVM algorithm is robust, accurate and very effective even in cases where the number of training samples is small.

Table 1.Comparative study of OCR Algorithms

OCR Algorithm	How it identifies the pattern of the character	Works for	Features
Template Matching Algorithm	By Comparing derived image features and templates	Only for typewritten characters	Easy to Implement, It only works on fonts of which it has templates.
Statistical Algorithm	By making observation and measurement	For typewritten and handwritten characters	It works even when prior data or information is not available about the character in the the training data.
Structural Algorithm	By Identifing the component of the character	For typewritten and handwritten characters	It uses structural shape pattern of the objects.
Neural Network Algoritm	By training Neural Network	For typewritten and handwritten characters	It also works when new fonts encountered, It is used to perform OCR due to their high noise tolerance.
Support Vector Machine	By mapping the input data onto a higher dimensional feature space and determine a separating hyperplane with maximum margin	For typewritten and handwritten characters	It is effective even in cases where the number of training samples is small; It achieves good generalization performance with no prior knowledge of the data.

IV. CONCLUSION

This paper presents study of OCR algorithms. How character patterns are identified by different OCR Algorithms in classification stage. Template matching algorithm is used for typewritten characters only. Structural algorithm, statistical algorithm, neural network algorithm and support vector machine is also used for handwritten characters. We can use any of the above algorithms as per our OCR program requirement.

REFERENCES

[1] Faisal Mohammad, Jyoti Anarase, Milan Shingote and Pratik Ghanwat, “Optical Character Recognition Implementation Using Pattern Matching”, International Journal of Computer Science and Information Technologies, Vol. 5(2), 2014.
[2] Mr. Danish Nadeem and Miss. Saleha Rizvi, “Character Recognition using Template Matching”, Department of computer Science, JMI.
[3] Rajib Lochan Das, Binod Kumar Prasad, Goutam Sanyal, "HMM based Offline Handwritten Writer Independent English Character Recognition using Global and Local Feature Extraction", International Journal of Computer Applications (0975 8887), Volume 46 No.10, pp. 45-50, May 2012.
[4] Pritpal Singh and Sumit Budhiraja, “Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script”. International Journal of Engineering Research and Applications (IJERA), Vol.1, ISSUE 4,pp.1736-1739.
[5] Jonathan J. HULL, Alan COMMIKE and Tin-Kam HO, “Multiple Algorithms for Handwritten Character Recognition”.
[6] Parveen Kumar, Nitin Sharma and Arun Rana, “Handwritten Character Recognition using Different Kernel based SVM Classifier and MLP Neural Network”, International Journal of Computer Science, Vol. 53, No.11, Sep.2012.
[7] Yusuf Perwej and Ashish Chaturvedi, “Neural Networks for Handwritten English Alphabet Recognition”, International Journal of Computer Application, Vol. 2, No. 7, April 2011.
[8] Md. Alamgir Badsha, Md. Akash Ali, Dr. Kaushik Deb and Md. Nuruzzaman Bhuiyan, “Handwritten Bangla Character Recognition Using Neural Network”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 11, Nov.2012.
[9] Dewi Nasien, Habibollah Haron & Siti Sophiayati Yuhaniz, Support Vector Machine (SVM) for English Handwritten Character Recognition, Second International Conference on Computer Engineering and Applications, 2010.
[10] Indrani Bhattacherjee, “Off-line English Character Recognition: A Comparative Survey”, Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing 2013.
[11] Anshuman Sharma, “Handwritten digit Recognition using Support Vector Machine”.
[12] Rohit Verma and Dr. Jahid Ali, “A-Survey of Feature Extraction and Classification Techniques in OCR Systems”, International Journal of Computer Applications & Information Technology, Vol. 1, Issue 3, November 2012.
[13] http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[14] Yasmine Elglaly, Francis Quek, “Isolated Handwritten Arabic Characters Recognition using Multilayer Perceptrons and K Nearest Neighbor Classifiers”.
[15] Rumiana Krasteva, “Bulgarian Hand-Printed Character Recognition Using Fuzzy C-Means Clustering”, Problems of engineering and robotics”, pp 112-117.
[16] http://link.springer.com/chapter/10.1007%2F978-3-642-05253-8_33#page-1
[17] Ganesh S Pawar and Sunil S Morade, “Realization of Hidden Markov Model for English Digit Recognition”, IJCA, Vol. 98– No.17, July 2014.
[18] Richa Goswami and O.P. Sharma, “A Review on Character Recognition Techniques”, IJCA, Vol. 83, No. 7, December 2013.
[19] Ms.M.Shalini, Dr.B.Indira, “Automatic Character Recognition of Indian Languages – A brief Survey”, IJISET, Vol. 1, Issue 2, April 2014.
[20] José C. Principe, Neil R. Euliano, Curt W. Lefebvre “Neural and Adaptive Systems: Fundamentals Through Simulations”, ISBN 0-471-35167-9
[21] http://www.heatonresearch.com/online/introduction-neural-networks-cs-edition-2/chapter-5/page4.html
[22] http://cdn.intechopen.com/pdfs-wm/40722.pdf

[23] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, L. Malik, M. Kundu and D. K. Basu, "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010
[24] C. J. C. Burges, .A tutorial on support vector machines for pattern recognition., Data Mining and Knowledge Discovery, 1998, pp 121-167.
[25] Antonio Carlos Gay Thomé, "SVM Classifiers – Concepts and Applications to Character Recognition".

AUTHORS

First Author – **Purna Vithlani**, Research Scholar, Department of computer science, Saurashtra University, Rajkot, India, e-mail: purna_vithlani@yahoo.com
Second Author – Dr. C. K. Kumbharana, Head, Department of computer science, Saurashtra University, Rajkot, India, e-mail: ckkumbharana@yahoo.com