Intelligent Character Recognition

College of Engineering, Pune

April 16, 2019

Shreyas Kulkarni	111608040
Venkatesh Yelnoorkar	111608077
Aditya Bhawalkar	141708001
Niramay Vaidya	111605075

Abstract

Character recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand in many emerging application domains, which require more advanced methodologies. As a subset of CR, special attention is given to the offline handwriting recognition, since this area requires more research to reach the ultimate goal of machine simulation of human reading. As a result, this project aims to create an offline handwritten character recognition model. A standardized process of achieving offline handwriting recognition which broadly consists of five steps, namely pre-processing, segmentation, representation, training and recognition and post processing. Pre-processing can be segmented into subtasks which consist of noise removal, binarization i.e. converting the image to grayscale, thinning, edge detection, slant estimation and correction, skew detection and resizing. Segmentation and representation will consist of dividing the cleaned image for feature extraction and then arranging the divisions into a standard format. Training and recognition will consist of the deep learning neural network architectural model which will involve a combination of convolutional and recurrent neural network layers. Post processing will consist of stitching the context of the information together to obtain meaningful text. Finally, a simplistic human interface will act as a interactive program enabling user to employ the proposed offline handwriting recognition application to digitize document images.

Introduction

Machine simulation of human functions has been a very challenging research field since the advent of digital computers. In some areas, which require certain amounts of intelligence, such as number crunching or chess playing, tremendous improvements are achieved. On the other hand, humans still outperform even the most powerful computers in the relatively routine functions such as vision. Machine simulation of human reading is one of these areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier.

In general, handwriting recognition is classified into two types as offline and online handwriting recognition methods. In the offline recognition, the writing is usually captured optically by a scanner and the completed writing is available as an image. But, in the online system the two dimensional coordinates of successive points are represented as function of time and the order of strokes made by the writer are also available. The online methods have been shown to be superior to their offline counterparts in recognizing handwritten characters due to the temporal information available with the former. However, in the offline systems, the neural networks have been successfully used to yield comparably high recognition accuracy levels. Several applications including mail sorting, bank processing, document reading and postal address recognition require offline handwriting recognition systems. As a result, the offline handwriting recognition continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy.

The first important step in any handwritten recognition system is pre-processing followed by segmentation and feature extraction. Pre-processing includes the steps that are required to shape the input image into a form suitable for segmentation. In segmentation, the input image is segmented into individual characters and then, each character is resized into m * n pixels towards the training network.

The selection of appropriate feature extraction method is probably the single most important factor in achieving high recognition performance. An artifial neural network as the backend is used for performing classification and recognition tasks. In the offline recognition system, neural networks have emerged as the fast and reliable tools for classification towards achieving high recognition accuracy.

Handwriting recognition is one of the many branches of computer vision, a vast field involving the mimicry of human capabilities of senses using massive technology advancements in the field of machine learning and aritificial intelligence. It mainly comprises of achieving machine vision to automate human-like tasks. Handwriting recognition has been an age old part of computer vision and with the advent of the technological boom in deep learning techniques, systems for this purpose have been successfully developed with astounding accuracy. Despite this, the offline handwriting recognition problem belonging to a broad category of recognition still remains unexplored and poses high research scope.

Literature Survey

Paper 1: Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network

In this paper, a diagonal feature extraction scheme for recognizing offline handwritten characters is proposed. In the feature extraction process, resized individual character of size 90 * 60 pixels is further divided into 54 equal zones, each of size 10 * 10 pixels. The features are extracted from the pixels of each zone by moving along their diagonals. This procedure is repeated for all the zones leading to extraction of 54 features for each character. These extracted features are used to train a feed forward back propagation neural network employed for performing classification and recognition tasks. Extensive simulation studies show that the recognition system using the diagonal features provides good recognition accuracy while requiring less time for training. The proposed recognition systems is given by-

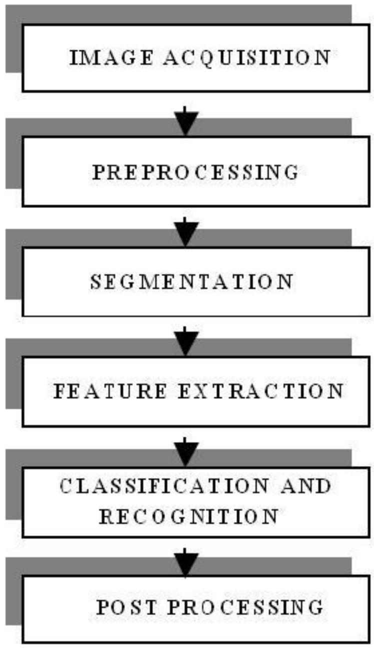

Figure 1: Schematic diagram of the proposed recognition system

Pre-processing consists of performing subtasks of noise removal, binarization, edge detection and dilation and filling on the scanned input image to further provide it to the feature extraction block. In the segmentation stage, an image of sequence of characters is decomposed into sub-images of individual character. The isolated characters are assigned a number for each distinct character as part of the labelling process. The labelling provides information about the number of characters in the image. Each individual character is uniformly resized into 90 * 60 pixels for classification and recognition stage. Then comes the diagonal feature extraction scheme as mentioned above. The features are extracted from each zone pixels by moving along the diagonals of its respective 10 * 10 pixels. Each zone has 19 diagonal lines and the foreground pixels present long each diagonal line is summed to get a single sub-featurem thus 19 sub-features are obtained from each zone. These 19 sub-features are averaged to form a single feature value and placed in the cooresponding zone. This procedure is sequentially repeated for all the zones. There could be some zones whose diagonals are empty of foreground pixels. The feature values corresponding to these zones are zero. Finally, 54 features are extracted for each character. In addition, 9 and 6 features are obtained by averaging the values placed in zones rowwise and columnwise, respectively. As result, every character is represented by 69, that is, 54 + 15 features. The classification stage is the decision making part of a recognition system and it uses the features extracted in the previous stage. A feed forward back propagation neural network having two hidden layers with architecture 54-100-100-38 is used to perform the classification. The hidden layers use the log sigmoid activation function, and the output layer is a competitive layer, as one of the characters is to be identified. The number of neurons in the hidden layers is chosen by trial and error. The neural network specifics are given by-

- Input nodes : 54 / 69
- Hidden nodes : 100 each

- Output nodes : 38 (26 alphabets, 10 numerals and 2 special symbols)
- Training algorithm : Gradient descent with momentum training and adaptive learning
- Perform function : Mean square error
- Training goal achieved : 0.000001
- Training epochs : 1000000
- Training momentum constant : 0.9

The comparison of recognition rate results obtained using different orientations with 54 features is given by-

Networks	1	2	3
Feature Extraction type	Vertical	Horizontal	Diagonal
Number of nodes in input layer	54	54	54
Number of nodes in 1st hidden layer	100	100	100
Number of nodes in 2nd hidden layer	100	100	100
Number of nodes in output layer	26	26	26
Recognition rate percentage	92.69	93.68	97.80

Paper 2: An Overview Of Character Recognition Focused On Off-line Handwriting

This paper focuses on the methodologies of CR systems, emphasizing the offline handwriting recognition problem. A heirarchical approach for most of the systems is given by ***pixel - feature - character - sub-word - word - meaningful text***. The paper gives an overall idea, a summary of various methods which can be used to perform various tasks present in the complete process of handwriting recognition. The pre-processing stage consists of noise reduction via filtering, morphological operations and noise modeling, normalization of the data via skew normalization and baseline extraction, slant normalization, size normalization and contour smoothing, compression in the amount of information to be retained via thresholding and thinning. The segmentation phase consists of external and internal segmentation which further consists of explicit and implicit segmentation and mixed strategies which combine these two techniques. Represention consists of global transformation and series expansion via fourier transforms, gabor transform, wavelets, moments and Karhunen-Loeve expansion, statistical representation via zoning, crossings and distances and projections, geometrical and topological representation via extraction of counting topological structures, measuring and approximating the geometrical properties, coding and graphs and trees. The training and recognition techniques enumerated are template matching via direct matching, deformable templates and elastic matching and relaxation matching, statistical techniques like parametric and non-parametric recognition, clustering analysis, hidden markov modeling and fuzzy set reasoning, strutural techniques like grammatical and graphical methods, and finally neural networks. Post processing consists of retaining contextual information of the entire text and in order to achieve this, a feedback loop is set up which feeds the information obtained after recognition back to the input of the system to improve the recognition model by reducing latency and increasing efficiency. Current studies in CR research is given by-

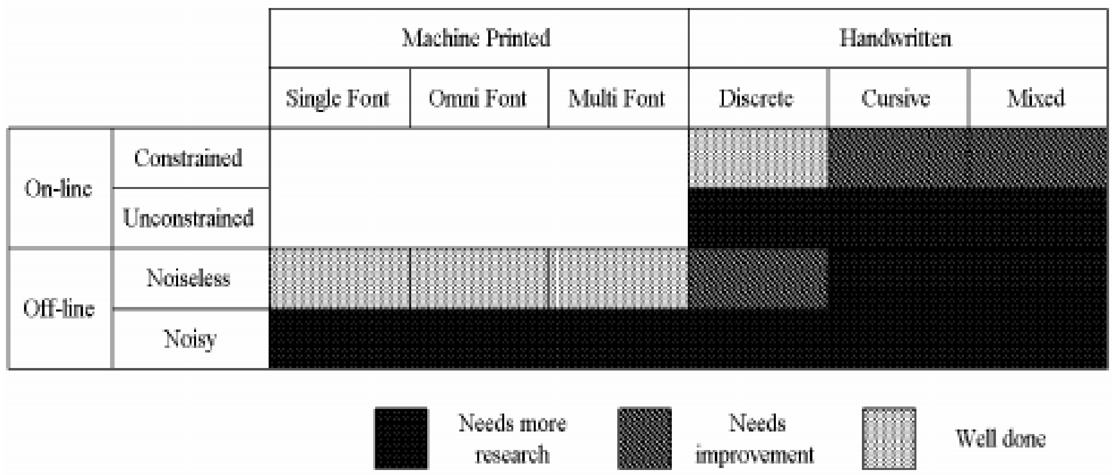

Figure 2: Current status of the existing intelligent character recognition systems

Paper 3: An Offline Approach to Handwriting Recognition

The paper mentions that the central tasks of offline handwriting recognition are character and word recognition. Document analysis is the necessary preliminary step in recognition that locates appropriate text when complex, two-dimensional spatial layouts are employed. According to this paper, pre-processing consists of thresholding, noise removal, line segmentation and word and character segmentation. This is given by-

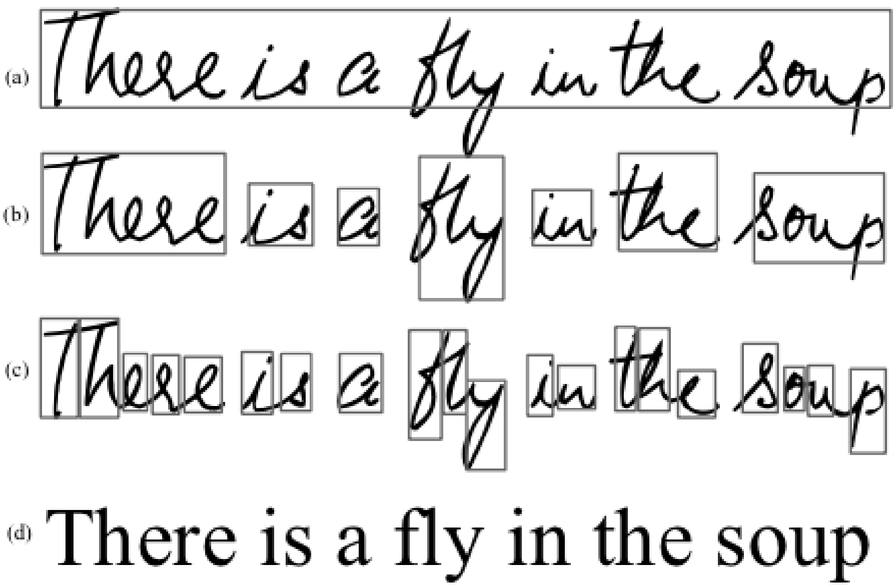

Figure 3: Line, word and character segmentation

The basic problem of character recognition is to assign the digitized character to its symbolic class. A pattern recognition algorithm is used to extract shape features and to assign the observed character to the appropriate class. Artifical neural networks have emerged as fast methods for implementing classifiers for OCR. In difficult cases, it becomes necessary to use models to constrain the choices at the character and word levels. Such models are essential in handwriting recognition due to the wide variability of hand printing and cursive script. A word recognition algorithm attempts to associate the word image to choices in a lexicon. Typically, a ranking is produced. This is done either by the analytical approach of recognizing the individual characters or by holistic approach of dealing with the entire word image. The latter approach is useful in the case of touching printer chacracters and handwriting. A high level of performance is observed by combining the results of both approaches. One method of word recognition based on determining pre segmentation points followed by determining an optimal path through a state transition diagram. Another approach utilizes the idea of regular and singular features. Handwriting is regarded as having a regular flow modified by occasional singular embellishments. A commmon approach is to use a Hidden Markov Model to structure the entire recognition process. Another method deals with a limited size dynamic lexicon. Words that are relevant during the recognition task are not available during training because they belong to an unknown subset of a very large lexicon. Word images are over segmented such that after the segmentation process no adjacent characters remain touching. Instead of passing on combinations of segments to a generic OCR, a lexicon is brought into play early in the process. A combination of adjacent segments is compared to only those character choices which are possible at the position in the word being considered. The approach can be viewed as a process of accounting for all the segments generated by a given lexicon entry. Lexicon entries are ordered according to the goodness of the match. This method is given by-

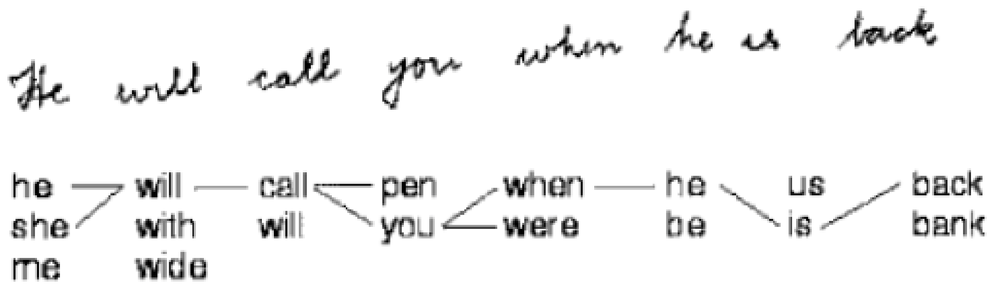

Figure 4: Recognition of a line

Dynamic programming (DP) is a commonly used paradigm to string the potential character into word candidates; some methods combine heuristics with DP to disqualify certain groups of primitive segments from being evaluated if they are too complex to represent a single character. The DP paradigm also takes into account compatibility between consecutive character candidates.

Paper 4: Offline Handwriting Recognition using Genetic Algorithm

In this paper, a new method for offline handwriting recognition is presented. A robust algorithm for handwriting segmentation has been described here with the help of which individual characters can be segmented from a word selected from a paragraph of handwritten text image which is given as input to the module. Then each of the segmented characters are converted into column in the form of text files. The network has been designed with quadruple layered neural network with 625 input and 26 output neurons each corresponding to a character from a-z, the outputs of all the four networks is fed into the genetic algorithm which has been developed using the concepts of correlation, with the help of this the overall network is optimized with the help of genetic algorithm thus providing us with recognized outputs with great efficiency of 71%. The handwriting recognition model described here works at three stages, segmentation of the handwritten text, recognition of segmented characters with the help of artificial neural networks and lastly selecting the best solution from the four artificial neural network outputs with the help of genetic algorithm.

The handwritten document is scanned and taken as an input to obtain individual characters which are written in a text file and are later read back and are passed as inputs to the four Artificial Neural Networks. In the algorithm, the scanned gray scale image is read into an image matrix which is converted into a monochromatic image matrix which pixel values of 0 for black points and 255 for white points. Then row wise searching is started from the point (0,0) to find out the first black point. This is the assumed top point of the first word of the handwritten text that has been inputted. This point is referred to the Upper Point. After the upper point is found, all the black pixels that are connected to this pixel are given a value of 999. Once this step is complete, then all the characters linked to that word have a value of 999 in the matrix under consideration. After finding all the connected points, a row wise search starts from the bottom to the top to find the first 999 value. This value corresponds to the Lowest Point. After this point is obtained, the area between the top and bottom point is searched on the left to check if any word has been missed on the left. In case another word is present on the left, then the new top point is obtained and the bottom point is found again else the procedure continues. After this is carried out, the left and right points of the word are found out by column wise searches. After the four points, the Upper, Lower, Right and Left point are found out, the word can be extracted and stored in a different matrix. This step is followed by marking of intersection points between various characters in the cursive handwriting. The word is searched for the number of cuts in a column wise manner. All the cases in which number of cuts is one and has an edge on either left or its right are marked and then all the successive markings are averaged to provide one optimum point through which a cut is marked with gray value of 0.5. After all the cuts are marked, in a loop all the characters are written into a file which is later read by the neural network to recognize the character.

The four artificial neural networks used consist of an input layer, three hidden layers and an output layer for each of the individual networks. The input layers takes the input from the image segmentation algorithm thus has 625 input neurons. The number of hidden layer neurons is as shown in the Table 1. The output layer consists of 26 neurons; this is due to the fact that there are 26 characters to be identified. Thus each output neuron corresponds to every character. Four artificial neural networks have been employed for the character recognition. The properties of each of the four artificial neural networks are designed using different parameters given by-

Name of Parameter		Artificial Neural Network 1	Artificial Neural Network 2	Artificial Neural Network 3	Artificial Neural Network 4
Number of neurons	Input Layer	625	625	625	625
	Hidden Layer1	8	8	6	8
	Hidden Layer2	16	14	16	12
	Hidden Layer3	8	8	6	8
	Output Layer	26	26	26	26
Training function(Algorithm)		trainlm (Levenberg-Marquardt algorithm)			
Number of Epochs		5000	4000	3000	2500
Performance Function		Mse (Mean square error)		Sse (Sum squared error)	
Training Goal		0.01	0.05		0.012
Memory Reduction Parameter		50			
Transfer Function		Purelin (Pure Linear)			

Figure 5: List of parameters used for the four artificial neural networks

A column vector is generated by the Image Segmentation consisting of 625 elements, which is saved in a .txt file. This .txt file is read and the elements are fed as an input to each of the four neural networks. These neural networks read all the weights and biases values that were saved in another files during the training process. Corresponding to the input, output is generated at the output layer. A 1 is set at the index of the characters that has been recognized. There can be more than one character that could be recognized based on the noise in the input of the neural network.

The following method is applied genetic algorithm to the outputs of the four artificial neural networks-

1. **Initialization** : Select the output of the neural network with the indexes comprising of "1's". This corresponds to the initial population for the genetic algorithm.

2. **Selection** : Select the indexes from the neural network that has minimum number of "1's".

3. **Fitness function** : Compute the correlation coefficients of the selected indexes.

4. **Mutation and crossover** : For the correlation coefficients less than the threshold value 0.50 repeat the step of the fitness function for a different training set. Discard the indexes that have coefficient values less than 0.3.

5. **Evaluation** : Select the index which has the maximum correlation coefficient with the input matrix.

6. Output the selected character.

Conclusion

The overall methodology of performing the task of offline handwriting recognition and developing a system prototype for it has been finalized by referring multiple researh papers which broadly delineate the common steps of pre-processing, segmentation, representation, training and recognition, and post processing. Various subtasks as part of these major tasks were studied and understood as mentioned by the referenced papers and a comparative study was performed to narrow down on a particular set of subtasks to be executed under these main tasks. This study is given by-

	Paper 1	**Paper 2**	**Paper 3**	**Paper 4**
Pre-processing	noise removal, binarization, edge detection, dilation and filling	noise reduction (filtering, morphological operations, noise modelling), normalization (skew, slant, size normalization, baseline extraction, countour smoothing), compression (thresholding, thinning)	thresholding, noise removal	image to matrix tranformation, conversion to monochromatic scheme
Segmentation	decomposition into sub-images and distinct character labelling	external, internal (explicit, implicit, mixed strategy)	line, word and character segmentation	word boundary detection, detection of join points between characters of a word
Representation	-	global transformation and series expansion (fourier transforms, gabor transform, wavelets, moments and Karhunen-Loeve expansion), statistical representation (zoning, crossings , distances, projections), geometrical and topological representation (extraction of counting topological structures, measuring and approximating the geometrical properties, coding, graphs and trees)	-	-

Training and Recognition	feature extraction using feed forward back propagation neural network	template matching (direct matching deformable templates, elastic and relaxation matching), statistical techniques (parametric and non-parametric recognition, clustering analysis, hidden markov modeling, fuzzy set reasoning), structural techniques (grammatical and graphical methods), neural networks	hidden markov model, limited size dynamic lexicon, dynamic programming	using combination of four neural networks with different training hyperparameters
Post processing	-	contextual information retainment using a feedback loop technique	-	employing a genetic algorithm to determine the best output out of the four distinct outputs generated by the four corresponding neural networks

References

1. International Journal of Computer Science and Technology (IJCSIT), Vol 3, No 1, Feb 2011

2. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216 - 233

3. Handwriting Recognition - Offline Approach, Department of CSE, Andhra University

4. Offline Handwriting Recognition using Genetic Algorithm, International Book Series Information Science and Computing

5. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000, "On-line and Off-line handwritten character recognition: A comprehensive survey"

6. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, July 1996, pp. 690-706, "A Survey of Methods and Strategies in Character Recognition"

7. Offline Handwritten Character Recognition Techniques using Neural Network: A Review, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064