Abstract

Conversion of handwritten text into digitized text and recognition of the same by a Computer is still a difficult and challenging task in today's much advanced hi-tech and materialistic world. Many different types of software, like advanced word processors, document scanning devices, sound recorders, voice recognizers etc., are available for versatile applications of text, speech etc. Still, it is difficult to find common software which can convert a handwritten text, written on a piece of paper, to its equivalent digitized form and recognizes it.

Difficulty arises in identifying the handwritten text because of many factors. The first problem arises when the characters are segmented and extracted out from the words. This is because of the fact that people use to write in different styles. Someone writes in cursive way where characters are connected. As a result, difficulty arises in finding the character boundaries. The situation becomes worsen when two characters are intermingled and appear as a single character. Lot of research work is going only on segmentation of characters.

The next difficult task arises at the time of preparation of input patterns that are presented to different character recognition systems i.e. at the time of preprocessing of the segmented characters. The most challenging task in recognizing handwritten characters is to recognize the variant patterns of same alphabet, which varies to a large extent even when written by the same individual at different instant of time. The final task is to compose the identified characters into digitized form.

On the basis of the above discussed issues, the work that has been carried out here focuses on development of some new handwritten character recognition models. These models can be used in a better way as compared to earlier systems in terms of efficiency as well as accuracy. The main objective is to extract out the common features of the varying patterns of handwritten characters and to develop methods that are capable of generalizing the variations in patterns to a great extent. Analyses of already developed handwritten character recognition systems are also explored.

This work proposes a method which is capable of identifying different texts taken from different individuals in different handwriting styles with great accuracy. All the

newly developed methods of this work are based on feature extraction techniques. Common features are extracted out in different ways, sometimes by segmenting the two dimensional character image matrices in different ways and sometimes forming unique codes for identical characters. Artificial Neural Networks (ANNs) are used to develop different handwritten text recognition models. This includes study of some classical models to identify various pros and cons in the system, identifying the gaps in the existing systems, developing some simple models by rectifying the problems of the previously developed models and finally developing some more improved versions where the performance is found much better than the previously developed methods.

Different work carried out at different stages in this work is divided into different phases which are discussed in different chapters of the thesis as follows:

Chapter-1 is the introduction of the overall work performed in order to develop some suitable handwritten character recognition model.

Chapter-2 is an overview of the Artificial Neural Network (ANN). Different types of ANNs are discussed briefly. Emphasis has been given on the recognition of handwritten texts using ANNs.

Chapter-3 is a detailed study of the literature related to the work. In this chapter, various pros and cons of the already developed methods and already carried out research works are discussed.

Chapter-4 is a short discussion of the software used to design the system. Texts taken from different individuals in handwritten forms are presented to test the performance of different existing and newly developed ANN models. MATLAB has been used as one of the main tools to design the handwritten character recognition systems. Different types of functions of MATLAB that are used for the purpose are briefly discussed here.

Chapter-5 is on a discussion and implementation of various preprocessing methods which are applied on the segmented characters to standardize the input patterns and also

to make the ANNs simple. Preprocessing of the input data is significant in the proper identification of the text data because it removes unwanted noise from the paragraph and simplifies the input vector.

There are various stages of preprocessing of the input data. This is done before and after segmentation of individual characters. Text paragraphs are scanned using a high definition scanner and saved in picture format. These paragraph images are converted into gray scale and then into binary matrices. The process of acquiring binary matrix is called binarization. Some additional spots of pixels may be present in the matrix as noise. These are removed by applying noise removal techniques.

Baselines are used for the segmentation of characters. A preprocessing technique called reference line estimation has been applied using baseline of the text.

Some post segmentation preprocessing has been applied in this work. The methods used are compression of character matrix, forming bipolar matrix and noise removal.

The number of elements in the vector represents the number of neurons in the input layer of the ANN. The binary matrix originally formed from the character image is actually very large in size which makes the ANN very large in size and complicated. In order to simplify the circuit the large two dimensional binary matrixes are compressed into small sized matrixes. A compression algorithm is also designed to compress the large sized matrix into a smaller sized matrix. The compression algorithm is designed in such a way so that the originality of the image matrix is preserved.

Chapter-6 is on development of different character segmentation methods. Sentences written in different styles by different individuals are presented to the developed Character Segmentation methods. Performances are measured for the developed methods.

Two character segmentation methods are developed to en-cage the characters extracted out from handwritten words and sentences. One of these is Slider Drifting Method (SDM) which easily identifies and extracts out the isolated characters present in words and sentences. The accuracy of SDM is 90%-100% for those handwriting styles where, there are gaps present between characters in sentences. But the performance of SDM is not good for the handwriting styles, where the characters are joined, especially for the cursive type of handwriting.

To overcome this problem another method has been developed, called Baseline Pixel Burst Method (BPBM), which is capable of identifying joined and cursive type of handwriting with a good accuracy. BPBM shows an average accuracy of 88%.

Chapter-7 is on studying and testing of some already developed traditional ANNs like Hebb and Perceptron training. A prototype model has been developed for the purpose. Two different forms of paragraphs are constructed using ten distinct handwritten alphabets. Individual characters are extracted out from the sentences of the paragraphs using the developed character segmentation method.

Segmented characters thus obtained are also preprocessed and presented to the developed prototype model of ANNs for training and testing and performances are noted. It is observed that performances of the classical models are very good for the printed alphabets but performances are not very much satisfactory for the handwritten characters. The gaps found in these ANNs are regarding the generalization of the varying patterns. An individual, when writes an identical alphabet twice at different instant of time fails to repeat the identical pattern for the same alphabet. Meager work has been done on these models to fill up this gap.

Chapter-8 describes the development of different single layer ANNs for the work. Single layer ANNs are developed to keep the ANN as simple as possible by keeping minimum number of neurons. The prototypes of the models are developed to test the sentences composed of only ten characters of English alphabet set.

The different models which are developed in this work are Row-wise Segmentation Technique (RST), Column-wise Segmentation of Image Matrix (CSIM), Input Pattern Segmentation Technique (IPST) and Row-wise Segmentation using Adaptive Resonance Theory ART1 (RSA). The performances of these single layer ANNs are found satisfactory. The models are based on different feature extraction techniques

To improve the performance, the work is carried out to the next stage by adding few additional layers.

Chapter-9 is on development of multiple layer ANNs. Multiple layer ANNs are formed by adding few extra layers in single layer ANNs. Extra layers enhance the performance of the ANNs in identifying the test samples which are deviated from their original position to a large extent. Actually, additional layers are used to put more weight on those input segments that contribute the most in forming the input patterns.

In a better developed model, additional layers are used to form unique codes for different characters. The multiple layer ANNs such as Arrow-Segmentation of Image Matrix (ASIM), Hoof Segmentation of Image Matrix (HSIM) and Pixel Density Gradient (PDG) Method are developed. It is found that the performance of the (PDG) Method is best among the other multiple layer ANNs.

Chapter-10 is on a comparative study. Performances (Accuracy) of all the methods developed here are compared with the other existing promising works found in the literature. It is established that the performance of the CSIM method is the best among the single layer ANNs and the PDG method is the best among the multiple layer ANNs.

Interfaces have been designed, using MATLAB software, for reading a handwritten paragraph and to segment it into different individual characters using different developed segmentation algorithms. The extracted handwritten characters are presented to different ANNs for training and testing. Finally, identified characters are reconstructed to words and sentences using MATLAB.