Abstract

The work in this thesis describes neural network based solution techniques for the selected critical problems in the automation of document processing.

Machine reading of optically scanned text of a document is usually called Optical Character Recognition (OCR), plays an important role in the document processing. Most of the existing work on OCR assumes that the language of the document to be processed is known beforehand, and the few available methods to identify the script or the language of document are so far primitive and are based on statistical techniques. The *Automatic Script Identification System* developed in this work to classify three languages English, Hindi, and Kannada is based on modular neural network architecture, which consists of three independently trained component feedforward neural network classifiers trained using a backpropagation algorithm. In this system, each document is represented by a set of five features extracted by newly developed *Hybrid Feature Extraction Technique.* This latest technique is the result of combination of a simple bar mask encoding methods and morphological operation on image called dilation. The unique feature of this system is, it classifies document images of size 64 x 64 pixels, whereas earlier systems required minimum size of 128 x 128 pixels. The system provides 99.0% accuracy on the indigenously created database which consists of 300 document images of size 64 x 64 pixels, a 100 each in three languages, English, Hindi and Kannada.

The *Modified Automatic Script Identification System* developed in the next level, overcomes the drawback of increased training time in the previous system. Additionally it works on an enhanced database of 1080 document images of the size 64 x 64 pixels, 120 in each of nine languages, English, Hindi, Kannada, Tamil, Gujarathi, Malayalam, Oriya, Telugu and Punjabi. An advanced recognition algorithms such as the radial basis neural network (RBNN) classifier and the probabilistic neural network (PNN) classifier, which form the central architecture of the system, are compared. The PNN classifier based system produces 97.4% overall classification accuracy.

The drawback of the above system is, it does not identify the script of individual word which is a major requirement in processing multi-script multi-lingual documents.

For this purpose, an *Individual Word Script Identification System* is developed. The system includes dynamic feature extractor, which can take care of variability in the length of an individual word document image, and PNN classifier. The system is designed with a training set of 225 words, 75 words in each of three languages English, Hindi and Kannada. The test set contains 225 unseen words, 75 words in each of above three languages. The system produces overall accuracy of 98.89%. The same system is compared with a system built with modular architecture that produces 98.00% accuracy.

A very important but largely ignored part in present neural based recognition systems that is, feature selection is taken up for further improvements. A *Feature Selection method using Genetic Algorithms* (GA) is developed. New methods to represent the problem and to design recognition rate based fitness function are presented. Three systems are built to recognize the script of a single document image, out of 1080 document images belonging to nine different script classes, and are compared. The first system which uses all the 50 extracted features yields 97.78% recognition rate, whereas second system which uses a GA engine selected 25 features produces 99.11% recognition rate. The third system that uses 25 features, which are selected based on the merit of individual performance, yields only 97.33%.

For further improvements in the above systems, three new *Hybrid Script Identification Systems* are presented based on three techniques of combing genetic algorithms and neural networks. The first genetic-neuro script identification system is a combination of hybrid feature extractor, GA based feature selector and a neural network classifier. The second neuro-genetic script identification system uses GA engine to optimize the parameter *spread* values of radial basis neurons of the PNN Classifier. Spread is responsible for the ability of recognition of the radial basis neuron. The system produces 98.7% accuracy with a 1080 document images database, containing 120 documents in each of nine script classes. The third modular genetic-neuro script identification system focuses on the concept of using class specific features for identification. The system developed to solve nine class script identification problem includes nine separate GA engines to select the best performing feature subset based on classification accuracy, out of 50 extracted features and nine PNN classifiers trained on

these corresponding class feature subsets. The results obtained on the same document image database used above, produces 99.63% overall classification accuracy.

Even though the above systems contribute to multi-script, multi-lingual document processing, there is a necessity for complete working model to process multi-scripted documents. The developed *Multi-script, Multi-lingual Document Processing* model is tested with a database of documents written in two languages. Each document contains six lines of text, out of which, three lines are written in English and three lines are written in Kannada script. The system so built, separates individual English word patterns and Kannada word patterns from such bilingual documents. The results of the experiments are very encouraging.

To implement the above developed multi-script, multi-lingual model there is a necessity of language specific character recognition systems. Hence a single font *Kannada Character Recognition System* is developed. It works on the concept of individual recognition of parts in four separate areas of each Kannada character, such as upper modifier area, core part area, subscript area, and horizontally extended area. The results of Kannada character recognition proved the effectiveness of the approach followed.

A *Bilingual OCR System* is built, embedding the above Kannada character recognition system and specifically developed single font English character recognition system. The system receives the given bilingual document written in English and Kannada languages. It separates words on the basis of language and generates two separate ASCII files.

In brief, the overall contributions of this research work include, (1) hybrid feature extraction technique, (2) document image script identification technique, (3) individual word script identification technique, (4) genetic based feature selection technique, (5) the three new hybrid systems for script identification, (6) multi-script, multi-lingual document processing model, (7) Kannada character recognition, and (8) the English and Kannada bilingual OCR system.