Chapter- 11

Conclusions

11.1 Conclusions

The work presented in this thesis describes neural network based techniques for document processing. Some of the critical problems in the automation of document processing and open research problems of neural based recognition systems are solved. The solution methodologies are based on artificial neural networks and newly developed hybrid document recognition techniques using combination of genetic algorithms and neural networks.

The approach adopted in this thesis is new because of the following reasons.

(i) In chapter 2, a new hybrid feature extraction technique has been presented. The benefits of morphological modification of an image, in the desired direction and the advantages of the bar mask encoding technique are successfully implemented in this newly developed hybrid technique. Comparisons with other techniques are also made and it has been proved that this technique provides very efficient features which are highly discriminative among different script classes of document patterns. This technique can also be extended to different recognition tasks such as character recognition, face recognition etc.

(ii) In chapter 3, an automatic script identification technique for document images belonging to three different classes, viz., English, Hindi and Kannada languages, is presented. Artificial neural networks are employed (in this thesis) for the first time to solve such a script identification problem. The modular neural network based system built is very efficient and also script-independent. It classifies 64 x 64 pixel sized document images where earlier systems failed. The same technique has been extended in chapter 4, as a script identification system for nine language documents (English, Hindi, Kannada, Tamil, Gujarati, Malayalam, Oriya, Punjabi and Telugu). The system uses a modified feature extraction technique which can extract 50 features for each pattern and probabilistic neural network classifier. The results obtained on a

1080 document image database of 64 x 64 pixels, shows that the system performs better than the radial basis neural network classifier. The detailed design procedures and methods to construct both probabilistic neural network and radial basis neural network are presented.

(iii) The very rarely attempted individual word script identification problem in multi-lingual documents is solved in chapter 5. The system presented uses the dynamic feature extraction technique to extract an equal number of features irrespective of size of input word document pattern. The results obtained on an indigenously developed database of 450 words, proves that the technique is very effective in classification of words of different scripts.

(iv) In chapter 6, genetic algorithms are used to select effective features for pattern recognition applications. This technique is very useful and saves measurement costs by eliminating the redundant features, which not only burden the recognizers but also add unnecessary complexities. A new representation scheme and new recognition rate-based fitness function are presented. After comparisons with systems built with all the extracted features and the features selected by best individual performances, it has been proved that the system built with GA-selected features provides the best results.

(v) Three new hybrid techniques of combining genetic algorithms and neural classifiers are illustrated in chapter 7. Corresponding to these three techniques three hybrid systems for script identification are presented. The first genetic-neuro hybrid system is the combination of GA-based feature selector and neural network classifier which are discussed in chapter 6. The second neuro-genetic hybrid system uses a probabilistic neural network classifier which has been designed using genetically selected spreads. This is a very new idea and the results obtained on the nine-class script identification system built using this technique, show that it can serve as an efficient pattern recognition model.

The third genetic-neuro modular script identification system is developed on the basis of the nature of human vision. The system uses genetically selected class-

specific feature subsets (similar to individual specific identification marks in personnel identification) and modular probabilistic neural network. The results obtained on 1080 document image database, 120 document images in each of nine script classes, are extremely good and prove that the developed model is very effective in solving critical pattern recognition problems.

(vi) In chapter 8, multi-script, multi-lingual document processing is addressed. The developed model presents a new method to process the documents printed in two or three languages. The results obtained on the database of bilingual documents printed in English and Kannada language scripts, prove that the model is very effective in classifying the individual words in such documents.

(vii) In chapter 9, the long standing unsolved problem of Kannada character recognition has been solved. New methodologies and algorithms for recognition of Kannada characters are presented. Document images of Kannada sentences are converted to ASCII files by the developed Kannada OCR system.

(viii) In chapter 10, for the first time, a bilingual OCR system for English and Kannada language documents is presented. Multi-script documents printed in these two languages are processed by the system. The system successfully separates the words based on the script and converted characters of these documents into separate ASCII files.

11.2 Summary

The gist of the contents and major contributions of the whole thesis, chapter-wise are presented in tables 11.1a and table 11.1b. These tables include the problem addressed in each chapter, existing methods, proposed methods in this thesis, datasets used, number of features used and specific contribution of that chapter. Summarizing the contributions of this research work: it presents, (1) hybrid feature extraction technique, (2) document image script identification technique, (3) individual word script identification technique, (4) genetic-based feature selection technique, (5) Three new hybrid techniques combining neural networks and genetic algorithms,

for script identification, (6) multi-script, multi-lingual document processing model, (7) Kannada character recognition, and (8) English and Kannada bilingual OCR system.

The last though very important factor in the conclusion is that all the work embodied in this thesis is done by keeping in mind the problems faced in the automation of Indian language document processing. Some of the problems solved here are very specific to Indian documents. Hence finally it can be concluded that this thesis has the genuine intention to contribute to Indian Society.

Table 11.1a: Overall summary sheet of the thesis.

Chapter Number ⟹	2	3	4	5
Problem addressed	Document representation-feature extraction	Script identification of document images	Script identification of document images	Individual word script identification
Existing state-of-art	Bar mask encoding moment invariants etc.	Statistical based script specific technique only useful for 128x128 pixel document images	Modular neural network based technique	Statistical based script separating technique only suitable for Bengali and Hindi
Thesis proposed state-of-art	Hybrid of bar mask encoding and morphological operation	Modular neural network based technique	Probabilistic neural network based technique	Probabilistic neural network based technique
Datasets used	15 document images of 64x64 pixels, belonging to 3 classes	300 document images, 100 each in English, Hindi and Kannada	1080 document images, 120 each in English, Hindi, Kannada, Tamil, Gujarati, Malayalam, Oriya, Punjabi and Telugu	450 word document images of various size, 150 each in three languages, English, Hindi and Kannada
Number of features used	- - -	5	50	50
Contribution/ Remarks	New feature extraction technique	Script independent technique, useful for even 64 x 64 pixels sized document images. 3 classes considered	Advantages are reduced design time, improvement in feature extraction, and better classifier. 9 classes considered	Dynamic feature extraction technique suitable for various sized document images. Solution to long standing problem of identification of script of individual word in multi-lingual documents

Table 11.1b: Overall summary sheet of the thesis (continuation of Table 11.1a).

Chapter ⟹ Number	6	7	8	9	10
Problem addressed	Feature selection in neural pattern recognizers	Document script recognition techniques	Multi-script, multi-lingual document processing model	Kannada character recognition	Bilingual document processing
Existing state-of-art	Statistical selectors, best individual features etc. (table 6.2)	Widely used technique is: Feature extractor + classifier, is widely used	----	(i)Available work concentrated on conveniently written small subset (ii)Presented recognized characters in English	Only bilingual OCR to read two languages Bengali and Hindi. Manual switching between languages
Thesis proposed state-of-art	Genetic algorithm-based feature selection technique	Three new techniques (1) GA engine selected features + classifier (2) GA engine selected spreads for radial basis neurons in probabilistic neural network (3) Class specific GA engines + modular network of specialized classifiers.	Neural based multi-script, multi-lingual document processing model	Prototype Kannada OCR system, which receives document image and converts to ASCII symbol set	Bilingual OCR system. It reads the bilingual documents and converts to two separate ASCII files
Datasets used	450 word document images, 150 each in 3 languages, English, Hindi and Kannada	1080 document images, belonging to nine language script classes	25 document images of 6 lines each (3 lines English and another 3 lines Kannada, in each document)	Document images written in commercially available Kannada font (one font). (may contain 2 to 6 lines)	Documents typed in two language scripts
Number of features used	Genetic engine selected 25 features	Technique dependent	50	----	----
Contribution/ remarks	New genetic-based feature selection technique	Three new hybrid techniques for recognition applications. Third technique simulates the human nature of vision.	Document segmentation techniques, method to process multi-script documents	Kannada character recognition system; converts document images to machine readable ASCII coded files	It is the first neural-based bilingual OCR system to read two languages. Completely automatic system