Text Recognition from PDF Files using BPNN and SVM

D.Sasirekha
Research Scholar
Karpagam University
Coimbatore, Tamilnadu

E.Chandra, PhD.
Director, Dept of Computer Science
Dr SNS Rajalakshmi college of Arts and Science
Coimbatore, Tamilnadu

ABSTRACT

OCR, is the process of electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. OCR systems are given additional consideration nowadays. The PDF files consist of text, images and graphs. Mixed Raster Content (MRC) technique segregates text and non-text region from the PDF files and the text part alone is extracted. Artificial Neural Networks (ANN) is a standard pattern classifier and extensively applicable to various problems and here uses Backpropagation learning algorithm which is very usable for image processing. SVM is a classifier that performs classification to find an optimal solution. Thus, this research uses the BPNN and SVM method for OCR from the extracted text files using features.100 different format of PDF files have been tested and the experimental results with recognition performance are tabulated by comparing both the techniques .

Keywords

MRC, OCR, ANN, BPNN, SVM

Abréviations:

PDF - Portable Document Format.

MRC - Mixed Raster Content.

OCR - Optical Character Recognition.

ANN - Artificial Neural Network.

BPNN - Back propagation Neural Network

SVM - Support Vector Machine

1. INTRODUCTION

Paper documents are important for transferring of information and communication. The conversion of paper document in a proper electronic form is essential for its processing, understanding, archiving, and transmitting by computers. The scope of this paper is the image processing of PDF Files, extracting the text part alone and recognizing for further process. For example, when a document is to be processed by an optical character recognition (OCR) system it is necessary to separate text from PDF files, so that time will not be wasted in attempting to interpret the graphics as text. . Character Recognition is a part of Pattern Recognition [1]. It is impossible to achieve 100% accuracy. The need for character recognition software has increased much since the outstanding growth of the Internet. Optical Character Recognition (OCR) is a very well-studied problem in the vast area of pattern recognition. The first commercial OCR systems began to appear in the early 1950s with the scanned images [2, 3].OCR programs are capable of reading printed text. This could be text that was scanned or converted images. Here by using MRC, the text part alone is extracted from the PDF images. Then the extracted image is then translated from an image format into a binary format, where each 0 and 1 represents an individual pixel. The binary data is then fed in to the neural network which has been trained for each characters and numerals. The characters could be of different size, orientation, thickness, format and dimension giving infinite variations. Recognition of characters is a challenging problem since there is a variation of the same character due the difference in font and sizes makes recognition task difficult, and also the recognition will not be accurate if the extracted text is not robust. Here we use back propagation neural network [4] that is used to recognize the characters. After training the network with back-propagation learning algorithm, high recognition accuracy can be achieved. SVM[16] is a training algorithm for learning with a strong theoretical foundation in statistical learning theory. Training data set was generated by labeling the features extracted from the test file to recognize a character.

This paper is organized as follows. PDF file described briefly in Section 2, Then the methodology is described in Section 3. In section 4, Experiments and results are discussed. Finally, we conclude in Section 5.

2. PORTABLE DOCUMENT FORMAT

PDF has been developed by Adobe for the distribution of electronic documents in a format that retains the exact look of the source material [5]. PDF files can be created by scanning a printed document or by using Adobe Acrobat (writer) to convert an electronic document, which has been produced by another application such as Microsoft Word, PageMaker or Quark XPress, into the Portable Document Format.

Adobe also provides Acrobat Reader, free software that allows PDF files to be viewed and printed using a variety of hardware and operating system platforms. The ability of PDF files to look exactly the same regardless of the system used to access them has led to the increasing use of PDF's on Websites in recent years. The Universal format PDF's are used largely due to non-editable, easy print and with secure password-protected documents.

3 METHODOLOGY

The methodology of this research work consist of 5 main levels containing PDF file to image conversion, Text Extraction ,Preprocessing, Feature extraction and finally classification is done using BPNN and SVM to recognize a character.

.

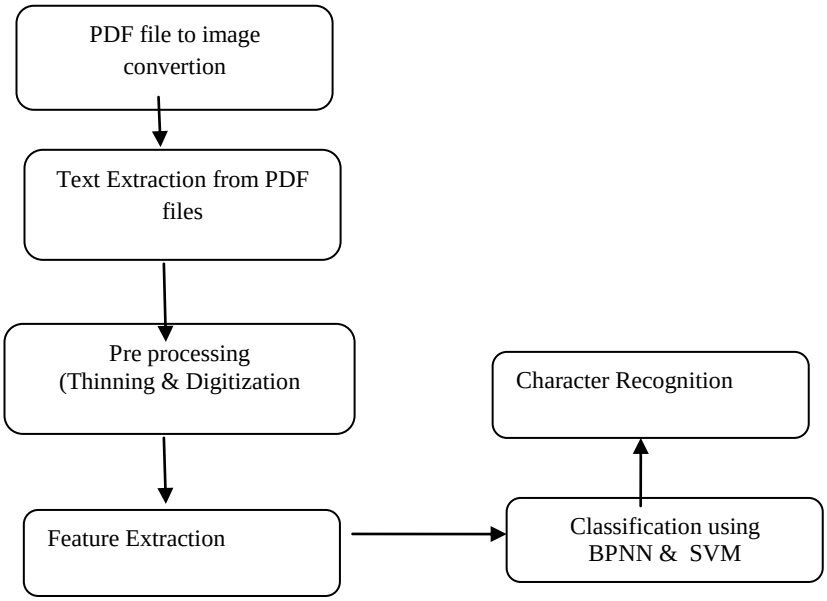

Fig 1: Character Recognition System

3.1 PDF Files to Images:

PDF files are converted into images using available commercial software's so that each PDF page is converted into image format. From that image format the text part are segmented and extracted for further process.

3.2 Text Extraction using MRC

Mixed Raster Content method is mainly used for compressing of compound images .Here takes a different approach. This method is modified slightly to suit the present work. The basic 3-layer MRC model represents a color image as two multi-level layers (Foreground or FG, and Background or BG) and one binary layer image (Mask or M). The mask layer describes how to reconstruct the final image from the FG/BG layers (Eq:1)

Image = Mask * BG + (1-M) * FG M {0,1} (1)

Depending on the mask value on a certain position, a pixel from the Foreground or Background on the corresponding position is selected (e.g. 0 for foreground selection and 1 for Background). Thus, the Foreground layer is poured through the mask onto the Background layer. An illustration of the imaging model is shown in Fig:2.

The basic 3-layer model is MRC's [11] most common form. The imaging model, however, is composed of basic elementary plane (layer) pairs, foreground and mask. With the mask layer the original image can be reconstructed again.

Here, By applying this method, the foreground (Text part) alone can be extracted for our next process.

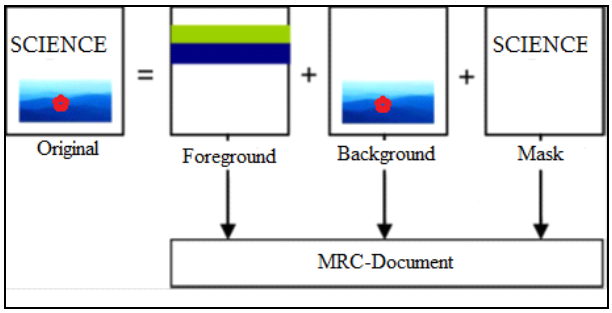

Fig 2 : MRC Model

3.2 Preprocessing (Thinning & Digitization)

Thinning, this operation is used to remove selected foreground pixels from binary images (Fig:3 (a)), so that the edges of the images are reduced to one pixel width. It is usually applied to a binary image and creates another binary image as output. The thinning of an image I by a structuring element J is given by

Thinning (I, J) = I – hit-and-miss (I, J)

Digitization transforms a binary image into a set of discrete points. The process of digitization is capturing corners where the lines end or change direction. The main benefit of digitization process is different sized images get converted to the same set of points (Fig.3(b)). This information is taken as input by the neural networks during learning stage.

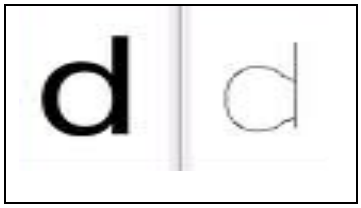

Fig 3: (a)Thinning process

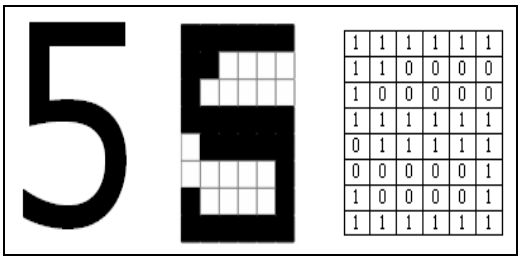

Fig 3 (b) Digitization Process

3.3 Feature Extraction

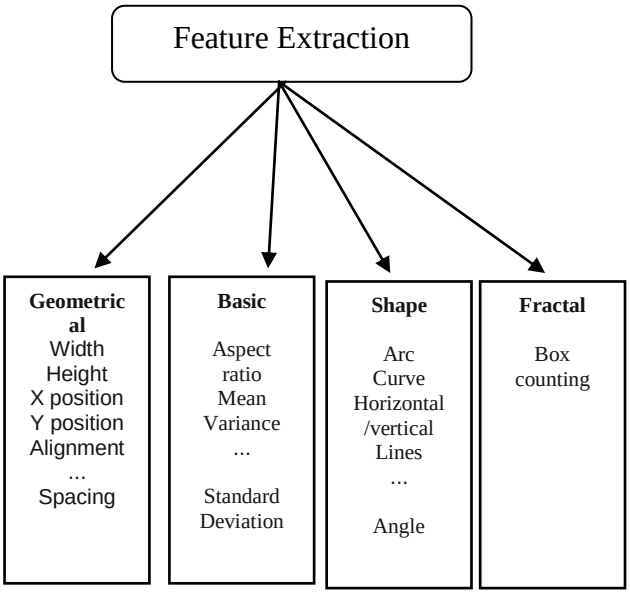

Fig 4: Feature Extraction

Research in printed document processing has been conducted for quite some time.Imadeet. al. [6] conducted segmentation and classification of printed characters, handwritten characters, photographs and painted image regions based on the histograms of gradient vector directions and luminance levels. Rule-based classification based on physical block structures such as width, height and ratio width to height has been used by Shih et. al. [7] which classify the segmented blocks into text, horizontal/vertical lines, graphics and pictures. Etemadet. al. [8] classifies the segmented blocks into image, text and graphics based on moments of wavelets. Graphics, photographs and text were classified in [9] based on color, texture and shape. Fractal dimension is computed using differential box counting method; the box-counting approach is one of the frequently used techniques to estimate the Fractal Dimension (FD) [10] of an image.

Using these features, 2 classifiers namely BPNN (Back Propagation Neural Network) and SVM (Support Vector Machine) are used to recognize the Character.

3.4 Classification

3.4.1 Back Propagation Neural Network

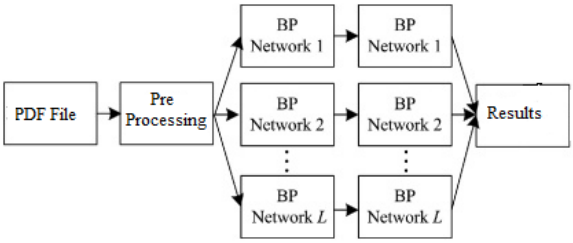

Fig 5: Back Propagation Neural Network Model

Back propagation neural network [12][13] is chosen for the classification purpose because it is simple to work and implement. The architecture of BPNN consists of three layers namely Input layer, Hidden layer and output layer. Where it consist of one input layer one hidden layer and one output layer for each character, recognized based on the feature. The features are extracted and passed through various layers and BPNN algorithm [20] [21] determines the output of each node, error node and corrected node. The performance is sensitive to the size of the character and it requires less storage due to less parameters. With limited iterations, errors are corrected.

3.4.2 Support Vector Machine (SVM) classifier

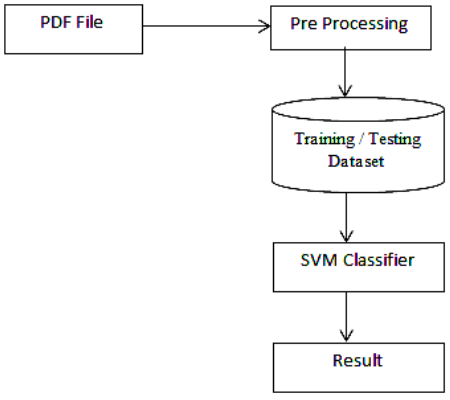

Fig 6: Supprt Vector Machine Model

The next classifier chosen for classification is Support Vector Machine (Fig:6). The reason for choosing SVM[14][15] is because of its popularity and for its outstanding recognition results in various pattern recognition applications and real world problems. SVM's classifiers based on Vapnik's [16][17][18][19] structural risk minimization principle

Table 1: Result Analysis of OCR System with BPNN and SVM

PDF file format	Metrices / Classifier	Basic Features		Geometric Features		Shape Features		Fractal Features	
		Error rate	Recognition rate	Error rate	Recognition rate	Error rate	Recognition rate	Error rate	Recognition rate
Single Column Test Files with no Figures	BPNN	0.0566	89.73	0.0522	91.17	0.0497	91.58	0.0542	90.38
	SVM	0.0549	92.81	0.0419	93.08	0.0368	93.91	0.0495	92.89
Double Column Test Files with no Figures	BPNN	0.0709	88.91	0.0637	90.25	0.0599	90.40	0.0683	89.54
	SVM	0.0670	89.11	0.0574	90.54	0.0521	91.25	0.0596	90.68
Single Column Test Files with Figures	BPNN	0.1075	89.07	0.0949	89.61	0.0897	89.88	0.0996	89.53
	SVM	0.0864	88.43	0.0724	90.61	0.0698	90.22	0.0784	91.23
Double Column Test Files with Figures	BPNN	0.1178	84.74	0.1036	86.31	0.097	87.14	0.1107	85.51
	SVM	0.0909	86.54	0.0863	87.85	0.0826	88.62	0.0884	87.06

which gives optimum results, fast training, avoids unnecessary complexity with a little prior knowledge or learning. During training SVM use large number of support vectors which must be stored and computed during the recognition of characters

4. Experimental Results:

The above discussed approaches are very robust in detection of individual characters. The Beauty of those approaches is that it is very simple to carry out. In this research back propagation neural network (BPNN) with input, hidden and output layer is used. In this experiment, extraction is done using 5 features (Fig:7). In SVM, a model is first created based on training samples. This model is then used to classify unknown data. The Goal of SVM is to find out a hyper plane with largest class margin, which best separate out given data.

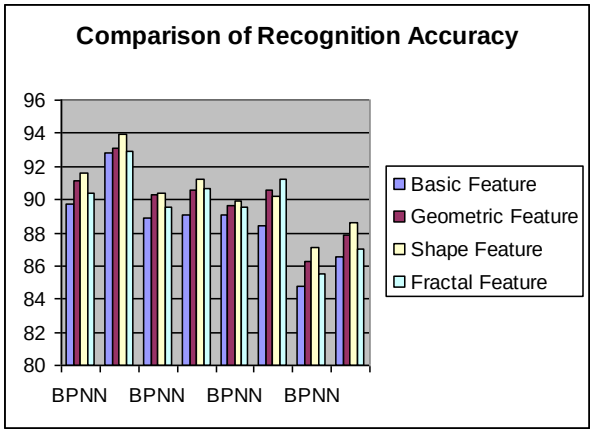

Fig 7: Comparison of Recognition accuracy based on different Features

5. CONCLUSIONS

This paper describes the simple and efficient OCR system for PDF files in English. Results have given quite satisfactory results for all the characters. It is quite faster as there is no complex processing involved. SVM has less chance of miss classification compared to neural network. From results, we can conclude that SVM out weights the performance of Neural network.

REFERENCES

[1] Andrew Blais and David Mertz, "*An Introduction to Neural Networks Pattern Learning with Back Propagation Algorithm*", Gnosis Software,Inc., July 2001.

[2] Yuelong Li Jinping and Li LiMeng, "*Character Recognition Based on Hierarchical RBF Neural Networks*", Intelligent Systems Design and Applications. Sixth International Conference, 2006, vol.1, pp. 127-132.

[3] Dong Xiao Ni Seidenberg, "*Application of Neural Networks to Character Recognition*", CSIS, Pace University, School of CSIS, Pace University, White Plains, NY, 2007.

[4] S. N. Sivanandam, S. N. Deepa," Principals of Soft Computing", Wiley-India, New Delhi, India.pp.71-83, 2008.

[5] Adobe Systems Incorporated, PDF Reference, Sixth edition, version 1.23 (30 MB), Nov 2006, p. 33.

[6] Imade, S.; Tatsuta, S. and Wada, T. "Segmentation and Classification for Mixed Text/Image Documents Using Neural Network", Proc. International Conference on Document Analysis & Recognition (ICDAR1993) , pp 930-934.

[7] Shih, F.Y. and Chan, S.S. "Adaptive Document Block Segmentation and Classification", IEEE Transactions on Systems, Man, and Cybernetics, vol. 26, no. 5, October 1996, pp. 797-802.

[8] Etemad, K,; Doerman, D.S. and Chellappa, R. "Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 1, pp. 92-96, Jan. 1997.

[9] Schettini, R.; Brambilla, C.; Ciocca, G.; Valsasna, A. And De Ponti, M. "A Hierarchical Classification Strategy For Digital Documents", Pattern Recognition 35 (2002), pp. 1759-1769.

[10] Jian Lia, QianDub,∗, CaixinSuna ,"An improved box-counting method for image fractal dimension estimation", Pattern Recognition 42 (2009) 2460 – 2469.

[11] EriHaneda, and Charles A. Bouman, "Text Segmentation for MRC Document Compression, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 20, NO. 6, JUNE 2011.

[12] Zaidah Ibrahim, Dino Isa, Rajprasad Rajkumar, Graham Kendall"Document Zone Content Classification for Technical Document Images Using Artificial Neural Networks and Support Vector Machines" 978-1-4244-4457-1/09/$25.00 ©2009 IEEE

[13] Gunvantsinh Gohil, Rekha Teraiya, Mahesh Goyani, "Chain Code And Holistic Features Based Ocr System For Printed Devanagari Script Using Ann And Svm", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.1, January 2012.

[14] Mamta Maloo et al. / International Journal on Computer Science and Engineering (IJCSE) "Support Vector Machine Based Gujarati Numeral Recognition", Vol. 3 No. 7 July 2011

[15] Arvind C.S., Nithya E And Nabanita Bhattacharjee " Kannada Language Ocr System Using Svm Classifier" Journal of Information Systems and Communication ISSN: 0976-8742, E-ISSN: 0976-8750, Volume 3, Issue 1, 2012, pp- 92-95.

[16] V.Vapnik, Statistical Learning Theory. John-Wiley and Sons , Inc., New York, 1998.

[17] Arvind C.S. Nithya E. And Nabanita Bhattacharjee3 "Kannada Language Ocr System Using Svm Classifier", Journal Of Information Systems And Communication, ISSN: 0976-8742, E- ISSN:: 0976-8750, Volume 3, Issue 1, 2012, Pp- 92-95.

[18] Htwe Pa Pa Win, Phyo Thu Thu Khine, Khin Nwe Ni Tun," Character Segmentation Scheme for OCR SystemFor Myanmar Printed Documents", International Journal of Computer Vision and Image Processing, 1(4), 50-58, October-December 2011.

[19] Bindu Philip ; R. D. Sudhaker Samuel," Preferred Computational Approaches for the Recognition of different Classes of Printed Malayalam Characters using Hierarchical SVM Classifiers" International Journal of Computer Applications, vol.I, Issue:16, Pg: 5- 10,2010

[20] Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal, "Survey of Methods for Character Recognition", International Journal of Engineering and Innovative Technology (IJEIT),Volume 1, Issue 5, May 2012.

[21] Madhup Shrivastava, Monika Sahu, Dr. M.A.Rizvi, " Artificial Neural Network Based Character Recognition Using Backpropagat", International Journal of Computers & Technology www.ijctonline.com ISSN: 2277-3061 Volume 3, No. 1, AUG, 2012