Document Segmentation And Region Classification Using Multilayer Perceptron

Priyadharshini N[1], Vijaya MS[2]

[1]Research Scholar, PSGR Krishnammal College for Women, Coimbatore, India

[2]Associate Professor, GR Govindarajulu School of Applied Computer Technology Coimbatore, India

Abstract

A document comprises lot of knowledge and documents are considered as the common mode of sharing information to others. Pursuance of information from documents involves lot of human effort, time consuming and can severely restrict the usage of information systems. Thus automatic information pursuance from the document has become a significant issue. It has been shown that document segmentation can help to overcome such issues. Document segmentation is a process of splitting the document into distinct regions. This paper proposes a new approach to segment and classify the document regions as text, image, graphics and table. Document image is segmented into blocks using Run length smearing algorithm and features are extracted from each blocks. Multilayer perceptron, a supervised learning technique has been used to construct the classifier and found 97.49% classification accuracy.

Keywords

Document analysis, Information retrieval, Classification, Feature extraction, Document segmentation.

1. Introduction

Document segmentation is defined as method of subdividing the document regions into text and non-text regions. A non-text region includes images, graphics, rules etc. Document segmentation plays a significant role in document analysis, because every day, millions of documents including technical reports, government files, Newspapers, books, magazines, letters, bank cheque, etc,. have to be processed. A lot of time, effort and money will be saved if it can be executed automatically. Automation of document analysis involves region extraction, identification of type of region and finally processing of each region separately. Document segmentation does the work of identifying the type of region. The documents contain both text and non-text regions. In order to process each region, the document should be segmented and then fed into the respective system for processing. For example, text regions are processed using OCR system. It converts text region into machine-readable form and non-text regions are conserved for processing such as compression, enhancement, recognition, and storage etc.

Some document segmentation applications are field extraction and recognition, word searching, logo detection, retrieving imaged documents in digital libraries, retrieval of documents containing tables or graphics. Also document segmentation is being adopted in postal industry where the address fields have to be identified before being sent to OCR readers and stamps have to be recognized. Generally document regions are segmented in two ways namely, geometric and logical based segmentation. In geometric based segmentation, the document is segmented upon its geometric structure such as text and non-text regions. Whereas in logical segmentation the document is segmented upon its logical labels assigned to each region of the document such as title, logo, footnote, caption, etc., [2].

Till now lots of methods have been proposed for document segmentation in the literature. Document segmentation techniques are broadly classified into three categories: top-down, bottom-up and hybrid approach [2]. A top-down approach repetitively segments the document image into smaller regions until further it cannot be segmented. Run-length smearing algorithm, projection profile methods, Fourier transforms etc., [4, 5, 7] are the methods which make use of top down approach. A bottom-up approach begins by merging pixels into characters. Then the characters are merged into words until whole document regions are merged. The methods which follow this approach are connected component analysis [3], run-length smoothing [6], region-growing methods [4], and neural networks [8]. A hybrid approach is the combination of both top down and bottom up approach. Few hybrid based methods are texture based and Gabor filters. The advantage of using top-down approach is, its high speed processing and the drawback is, it cannot process table, improper layout documents and forms.

This research work associates the existing features specified in [4] [6] [8] and proposes few features which subsidizes more in document segmentation. Features such as perimeter/height ratio, energy, entropy are employed. Perimeter/height ratio is defined as fraction perimeter to height of the block. A block in document image is a connected component and it is defined as a collection of black runs that are 8-connected.Both perimeter and height of the block diverges in their values. Text blocks have slighter value for perimeter/height ratio when compared to non-text blocks. Thus perimeter/height ratio is essential in classifying the blocks.

Additionally energy and entropy features are used to classify the blocks. Energy and entropy are renowned properties of an image. Energy identifies the uniformity of the image. Whereas entropy identifies the randomness (texture) of the image. Each block of the document varies in its energy and entropy specifically in case of table, graphics and image blocks. Thus these new features offer a notable influence in document segmentation.

This paper presents the implementation of bottom up approach to segment the document image into homogenous regions based on geometric structure. As bottom up approach is well suited for different layout documents and documents with graphics, image etc., document image is segmented into blocks using run length algorithm. The segmented blocks are then classified into the following categories: text, image, graphics, table using multilayer perceptron and implemented in WEKA.

2. System Overview

The proposed approach reduces the computational complexities of supporting procedure after document segmentation such as OCR system for processing text regions and for processing non text region. It reformulates document segmentation problem as classification task and solved using Multilayer perceptron. Documents are collected from various journals and features are extracted from document images. The proposed method consists of four phases: preprocessing, segmentation using Run length smearing algorithm (RLSA), labeling connected components and feature extraction. Each phase is described in the following sections and system architecture is shown in Fig. 1.

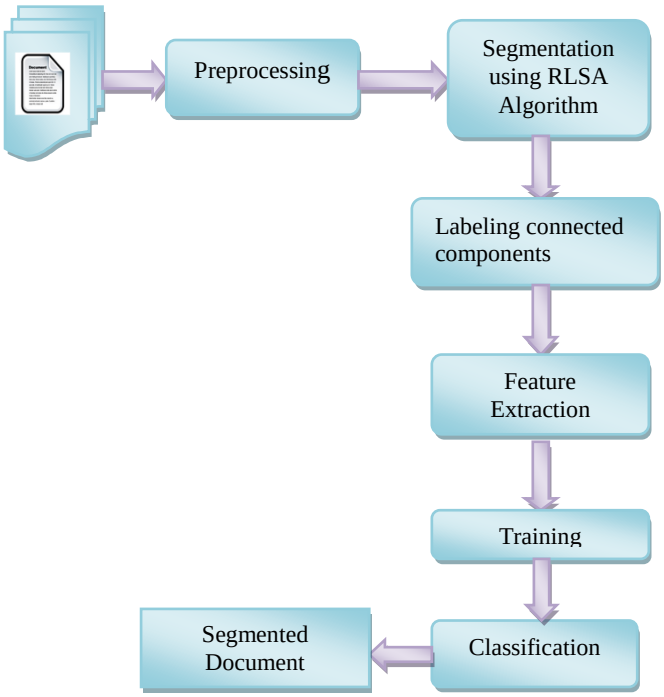

Fig. 1 .System Overview

2.1 Preprocessing

Preprocessing is a sequence of tasks performed on the document image. It enhances the quality of the image for segmentation. The various tasks performed on the image in preprocessing stage are scanning, binarization, and noise removal.

Scanning:

The documents are collected from various journals and scanned at 200 dots per inch (dpi).

Binarization:

It is a process which converts the grayscale image into a binary image using the global threshold method. A binary image has only two values 0 or 1for each pixel. 0 represents white pixel and 1 represents black.

Noise removal:

From the binarized image, the noise is removed using ostu method. The obtained noiseless image is subjected to RLSA algorithm for further processing.

2.2 Segmentation using RLSA algorithm

The RLSA algorithm is used for segmenting the document. The document is subdivided into blocks, where each block contains only one type of data such as text, graphic, halftone image, etc.,.

The Run length algorithm is employed to a binary image where 0 represents white pixels and 1 represents black pixels. The algorithm transforms a binary image **x** into an output image **y** as follows:

1. White pixels in x are replaced with black pixels in y if white runs are less than or equal to a predefined threshold.

2. Black pixels in x are untouched in y.

With a selection of optimal threshold value, the connected areas will form blocks of the same region. The Run Length Algorithm is employed horizontally with horizontal threshold hTh = 300 as well as vertically with vertical threshold vTh = 280 to a document image, producing two individual images. The two images are then combined using logical AND operation. Furthermore, horizontal smearing is applied with horizontal threshold Th = 30 to generate segmented blocks. Fig. 2 shows the segmentation of document image using RLSA algorithm.

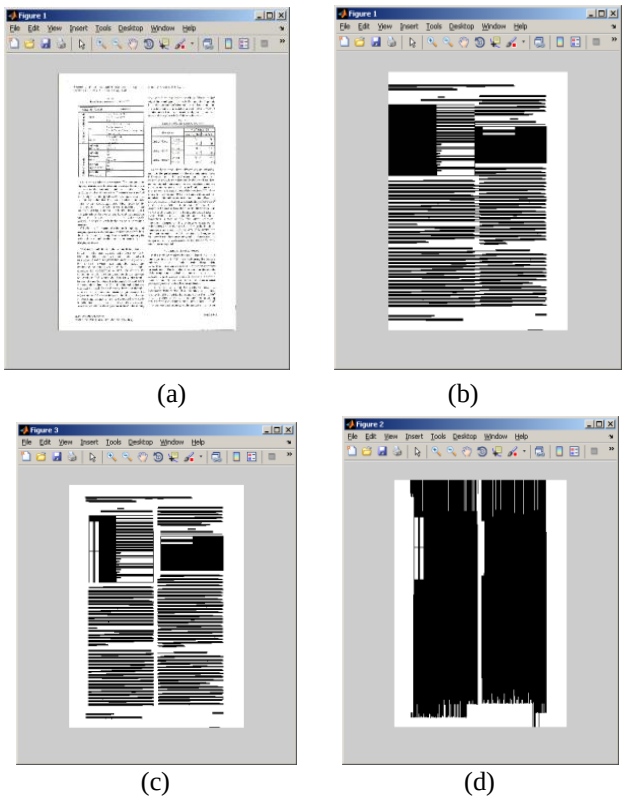

Fig. 2. (a) Original document image. (b) Result of horizontal smearing. (c) Result of vertical smearing. (d) Final result of segmentation.

2.3 Labeling connected components

Labeling is the process of identifying the connected components in an image and assigning each component a unique label an integer number which must be same as connected black runs. Labels are used to differentiate each block. Fig. 3 shows the labeled connected components of document image.

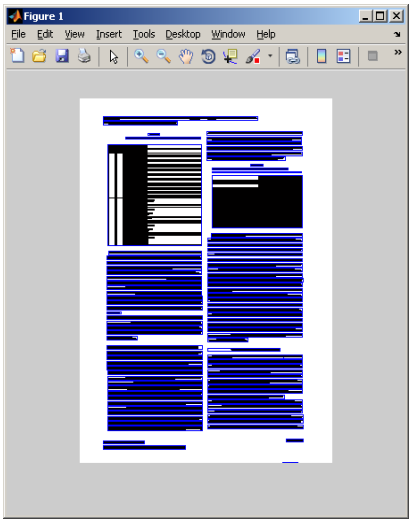

Fig. 3.Labeling Connected components

2.4 Feature Extraction

Feature extraction is the process of transforming the document image block into set of features. The features describing the properties of each block are extracted based on its edges. The coordinates of the edges, specify the size of the block. Each extreme edges of block are defined by considering the origin of binary image.

Features such as height, area, aspect ratio, perimeter, perimeter/height ratio, average horizontal length are extracted to differentiate the non-text block from text block. Image, graphics and table blocks are identified by means of following features: density, density of segmented block, horizontal transition along x-axis and y-axis, vertical transition along x-axis and y-axis. The table and graphics blocks can be recognized more accurately using the features such as mean standard deviation, active pixels, energy and entropy.

Blocks are represented by (Xmin, Ymin) and (Xmax, Ymax).Where,

Ymin - Left-most pixel value of column.
Ymax - Right-most pixel value of column.
Xmin- Left-most pixel value of row.
Xmax - Right-most pixel value of row.

The following block features are computed for classification of document region.

Height (H) - Height of the block is computed by subtracting the leftmost pixel from rightmost pixel of column.

$$H=Dx = (Ymax-Ymin) +1$$

Width (W) - Width of the block is computed by subtracting the leftmost pixel from rightmost pixel of row.

$$W=Dy = (Xmax-Xmin) +1$$

Aspect Ratio (E) - It is defined as the ratio of a block's width-to-height.

$$E=W/H$$

Area (A) - Area of the block is obtained by multiplying the height and width.

$$A=H*W$$

Density (D) - Density is defined as the ratio of total number of black pixels within each block of document image to the area.

$$D=N/Area$$

Horizontal transition along x axis (HTx) - It is defined as ratio of horizontal transitions per unit height and computed as

$$HTx=HT/Dx$$

Where, HT is the horizontal transitions of black to white or white to black pixels in a block of document image.

Vertical transition along x axis (VTx) - It is defined as ratio of vertical transition per unit height and is computed using the following formula.

$$VTx=VT/Dx$$

Where, VT is the vertical transitions of black to white or white to black pixels in a block of document images.

Horizontal transition along y axis (HTy) - It is defined as horizontal transition of white to black pixels per unit width and is given by

$$HTy=HT/Dy$$

Vertical transition along y axis (VTy) - It is defined as vertical transition of white to black pixels per unit width and is given by

$$VTy=VT/Dy$$

Density of segmented block (D1) - It is the ratio of total number of black pixels of segmented block after applying Run length smearing algorithm to the area.

$$D1=C/Area$$

Average Horizontal Length(R) - It is defined as mean horizontal length of black runs of document image within each block.

$$R =N/HT$$

The Product of block height and ratio R is given by

$$RH=R*H$$

The Product of ratio R and ratio E is given by

$$RE=R*E$$

The Product of ratio D1 and ratio R is given by

$$RD=R*D1$$

Mean (E) - Mean value of pixel intensities in the block is defined as arithmetic average of distribution of pixels in the block. It is computed using the function mean () in matlab.

$$E_i^{n-1} = \sum x_i / N$$

Standard Deviation (σ) - Standard deviation of pixel intensities in the block is defined as calculated using the function std () in matlab.

$$\sigma_i^{n-1} = \sum (x-\mu)$$

Active pixels (A) - Active pixels of the block is defined as the count of pixels with intensity < mean-k*std and it is computed as follows.

$$A = Sum (I< (mean-k*std))$$

Perimeter (P) - Perimeter of the block is defined as the distance around the block and computed using the formula

$$P = 2*(H+W)$$

Ratio S - It is defined as ratio of perimeter to height of the block and is given by

$$S=Perimeter/height$$

Energy (E) - It is defined as sum of squared elements in the image and is also known as uniformity or the angular second moment.

$$E_{i=1}^{n} = \sum x_i^2$$

Entropy- It is defined as statistical measure of randomness that can be used to characterize the texture of the input image. It is calculated using the function entropy () in matlab.

$$\text{Entropy} = \sum_{i,j} P(i,j) \log P(i,j)$$

Thus a group of 20 distinct features are extracted from the blocks of document image using matlab. The feature vectors are generated using 519 segmented blocks of 15 document pages and the training dataset is generated with 519 instances.

3. Multilayer Perceptron – A Neural Network Model

Neural network is a supervised learning algorithm. Supervised learning is the machine-learning task of inferring a function from supervised training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value called the supervisory signal. A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The classifier is then used for predicting the accurate output value for any valid unseen input object.

An Artificial Neural Network is a computational model that is inspired by the biological nervous system. The biological neuron consists of four components namely dendrites, soma, axon and synapses. Dendrites form a hair-like set of extensions of the soma. They receive signals (information) from other neurons. Synapses modulate the signals and forward the signals to axon. Axon conducts signals away from the cell body to other neurons. Fig. 4 shows the biological neuron structure.

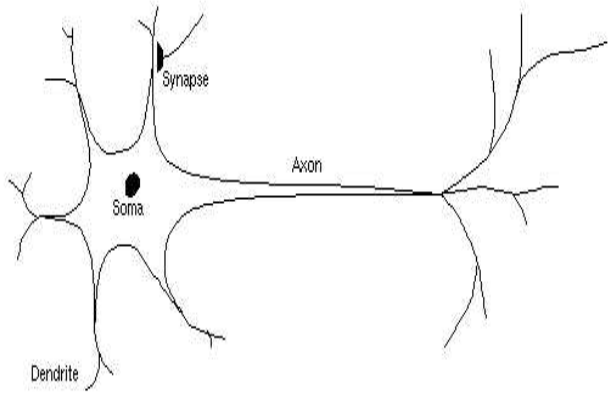

Fig. 4. Biological neuron

The artificial neuron inspired from biological neuron is represented in Fig. 5.

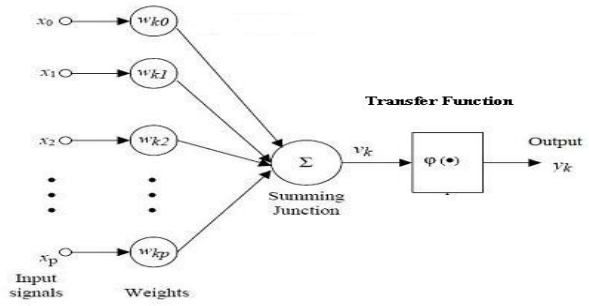

Fig. 5. Artificial neuron (perceptron)

In artificial neuron, dendrites act as input channels which receive the features as input. The inputs are multiplied by weight as synapses modulate in biological neurons. The product is summed and fed into the transfer function (axon) to produce the output.

Multilayer Perceptron (MLP) network is the most widely used neural network classifier. A multilayer perceptron is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable by a hyper plane. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Multilayer perceptron using a back propagation algorithm are the standard algorithm for any supervised-learning pattern recognition process.

4. Experiment and results

The document image classification model is generated by implementing supervised learning algorithm. The documents used for creating the dataset are collected from various journals and here 15 document pages have been used. The dataset consists of 519 segmented blocks of which 196 are text blocks, 123 are graphics blocks, 92 are image blocks and 108 are tabular blocks. The features describing the properties of the blocks are extracted and the size of each feature vector is 20. The dataset with 519 instances is trained with multilayer perceptron using WEKA. The Weka, an Open Source, Portable, GUI-based workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools [17].

The experiment is also carried out using other classification algorithm such as decision tree induction and naïve bayes. The same dataset have been used for learning and implementing J48 and naïve Bayes algorithm.The classifiers are built using WEKA.

The performance of the classifiers is evaluated using 10–fold cross validation and comparative analysis has been carried out. Classification accuracy is used as a primary performance measure for evaluating the classifiers and is measured as the ratio of the number of correctly classified instances in the test dataset and the total number of test cases. The performances of the trained models are evaluated based on two criteria, the classification accuracy and the training time. The results of three models in terms of prediction accuracy and time taken is shown in Table 1 and the comparative performance of classifiers is depicted in Fig 6.

Table 1 Performance Comparison Of Classifiers

Evaluation criteria	Classifiers		
	J48	***NB***	***MLP***
Time taken to build model (Sec)	0.04	0.01	4.32
Correctly classified instances	499	490	506
Incorrectly classified instances	20	29	13
Prediction accuracy	96.14%	94.41%	97.49%

Table 2. Comparison of estimates

Evaluation criteria	Classifiers		
	J48	***NB***	***MLP***
Kappa statistic	0.7469	0.7129	0.8692
Mean absolute error	0.0202	0.0287	0.0143
Root mean squared error	0.1369	0.165	0.0939
Relative absolute error	25.0854 %	35.6537 %	17.7152 %
Root relative squared error	69.358 %	83.5725 %	47.5704 %

Table 3. Confusion matrix of the classifiers

Classifiers	Text	Table	Graphics	Image
J48	194	0	2	0
	0	87	8	0
	6	0	108	2
	0	0	2	92
NB	181	0	6	3
	0	108	0	0
	0	1	109	8
	0	1	10	65
MLP	195	0	1	0
	1	109	0	0
	9	2	103	0
	0	0	0	96

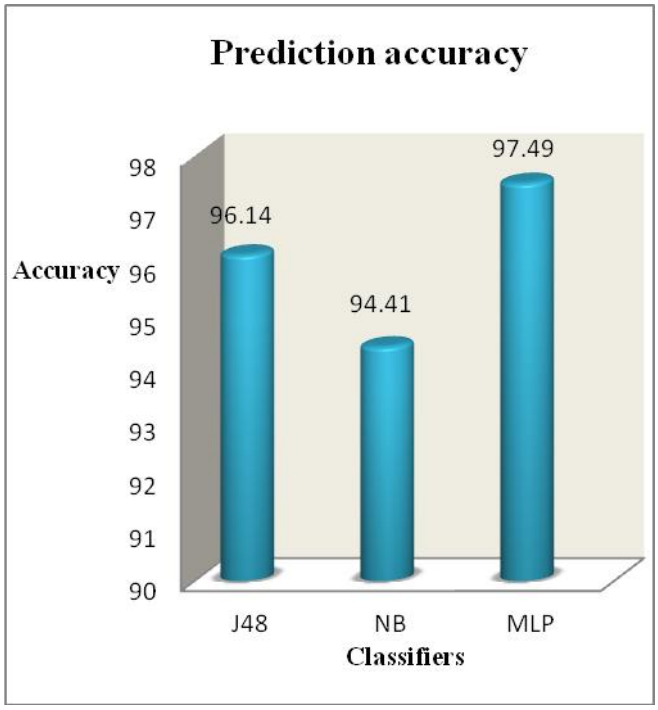

Fig. 6. Prediction accuracy of classifiers

Fig. 6 illustrates the comparative results of the three algorithms. Multilayer perceptron performs well than other two algorithms. Among the three classifiers used for the experiment, the decision tree induction algorithm (J48) and Multilayer perceptron algorithm provides more or less the same prediction accuracy. The accuracy rate of Naïve Bayes classifier is the lowest among the three machine learning techniques.

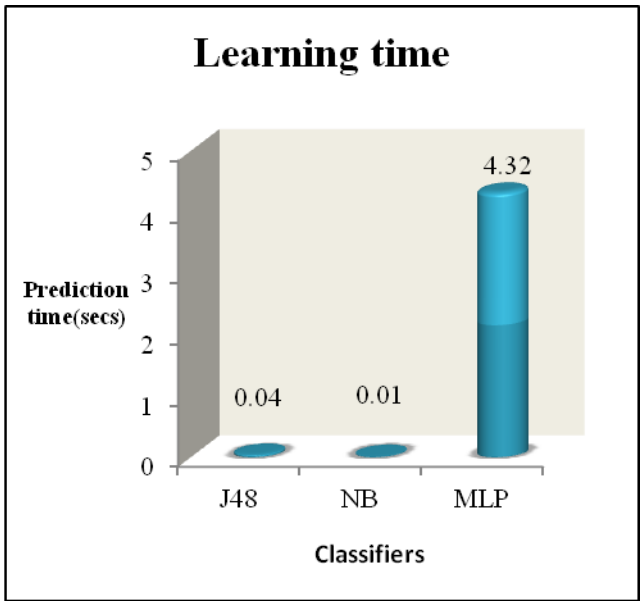

Fig. 7. Learning time of classifiers

From the above comparative analysis it is observed that classification accuracy produced by multilayer perceptron is higher than naïve bayes and J48. The time taken to build the model is high in case of multilayer perceptron when compared to other two algorithms. As the document segmentation is adopted in areas like information retrieval, the built model can be incorporated into such IR systems and hence the prediction accuracy plays major role in evaluating the performance of document segmentation model than learning time. Hence it is concluded that multilayer perceptron performs well than J48 and naïve bayes.

5. Conclusion

This paper demonstrates the modeling of document segmentation as classification task and describes the implementation of machine learning approach for segmenting the document into various regions. J48, Naïve Bayes, Multilayer Perceptron have been applied for generating classification models. The performance of the models has been evaluated using, classification accuracy and learning time. From the results it has been found that Multilayer Perceptron performs better than decision tree and Naive Bayes. Future task is to segment and recognize the handwritten document.

References

[1] Rangachar Kasturi, Lawrence O'gorman and Venu govindaraju. "Document Image Analysis: A Primer".

[2] O. Okun, D. Doermann and M. Pietikainen. "Page segmentation and zone classification". The state of the art. In UMD, 1999.

[3] Jain A and Yu B. "Document representation and its application to page decomposition". IEEE trans. On Pattern Analysis and Machine Intelligence, 20 (3): 294–308, March 1998.

[4] Frank Y shih,shy-shyan chen, Douglas Hung DC and Peter A Ng. "A Document segmentation, classification and recognition system"

[5] Wahl F, Wong K, and Casey R. "Block segmentation and text extraction in mixed text/image documents". CGIP, 20:375–390, 1982.

[6] Wang D and Srihari S. "Classification of newspaper image blocks using texture analysis". CVGIP, 47:327–352, 1989.

[7] Pavlidis T and Zhou J. "Page segmentation by white streams". Proceedings 1st International Conference Document Analysis and Recognition (ICDAR), International journal of Pattern Recognition, pages 945–953, 1991.

[8] Tan C and Zhang Z. "Text block segmentation using pyramid structure". SPIE Document Recognition and Retrieval, San Jose, USA, 8:297–306, January 24-25 2001.

[9] Le D X, Thomas G, Weschler H. "Classification of Binary Document Images into Texture or Non-textual Data Blocks Using Neural Network Models".

[10] Hose M and Hoshino Y. "Segmentation method of document images by two-dimensional Fourier transformation". System and Computers in Japan.

[11] Jain AK and Bhattacharjee S. "Text segmentation using Gabor filters for automatic document processing". Machine Vision and Applications, 5 (3): 169–184, 1992.

[12] O'Gorman, Kasturi L. "Document Image Analysis". IEEE Computer Society Press, Los Alamitos, California, 1995.

[13] Ha T, Bunke H. "Image processing methods for document image analysis". Handbook of character recognition and document image analysis. World Scientific 1–47, (1997).

[14] Kise k, Sato A and Iwata M. "Segmentation of page Images using the area Voronoi Diagram". Computer Vision and Image Understanding 70: 370-382, 1998.

[15] Machine Learning Tools and Techniques. Elsevier Gupta GK "Introduction to Data Mining with Case Studies".

N. Priya dharshini is pursing Master of Philosophy in Computer Science in PSGR Krishnammal college for women under the guidance of MS.Vijaya. Her research interests are data mining, image processing, and pattern recognition.

MS. Vijaya is presently working as Associate Professor in GR Govindarajulu School Of Applied Computer Technology, PSGR Krishnammal college for women, Coimbatore, India. She has 22 years of teaching experience and 8 years of research experience. She has completed her doctoral programme in the area of Natural Language Processing. Her areas of interest include Data Mining, Support Vector Machine, Machine learning, Pattern Recognition, Natural Language Processing and Optimization Techniques. She has presented 22 papers in National conferences and she has to her credit 17 publications in International conference proceedings and Journals.
She is a member of Computer Society of India, International Association of Engineers (Hong Kong), International Association of Computer Science and Information Technology (IACSIT – Singapore). She is also a reviewer of International Journal of Computer Science and Information Security.