Classification of Binary Document Images Using Probabilistic Neural Network Model

Eman H. Abdurrahman
College of Education for Girls , University of Mosul , Mosul , Iraq
(**Received:** 17 / 1 / 2010 ---- **Accepted:** 2 / 4 / 2012)

Abstract
In this work, a proposed system is used to segment and classify the gray document image to two regions texture and non-textual based on data blocks. A probabilistic neural network model has been used for this purpose . First the preprocessing is used to convert gray document image to binary document image, then the erode process used to convert the binary document into blocks. The resulting blocks are segmented to number of regions by using label procedure, the four features of each region are calculated based on the bounding box for each region, Then these features are fed to the input layer of a probabilistic neural network for classification to one of two regions are text, non-textual. Some gray documents images are used in order to test the proposed system. Finally three experiments are applied and the algorithm classified them correctly

Keywords: Binary image classification, probabilistic neural network.

1. Introduction
Todays' computers equipped with cameras or optical scanners can read documents and provide their faithful electronic reproduction. In spite of these technological achievements, however, stacks of documents still flood desks of most offices. Whereas documents can be read and accurately stored, the processing required for extracting information is still only in its infancy. The automatic information extraction is very hard because of the noise or the hardly predictable document structure. In the last years, the decreasing cost of document collecting, storage, processing, and the interest in Artificial Neural Networks (ANNs) have given rise to many novel solutions to different tasks of document processing.

This paper describes the classification of gray document images into blocks of textual and nontextual data using probabilistic neural network (PNN) model. It used because of its ability to estimate probability density functions (pdfs) based on training is instantaneous, it is not hard to adapt to new incoming pattern. Textual data usually consists of letters, words, and/or sentences, while examples of nontextual data include graphics, forms, and line art.

The proposed classification algorithm based on first extracting features for unknown gray document image. The second is using these features as input into neural network to classify it based on these features into these regions. Documents classification, as related to image analysis involves taking the features extracted from the document image and using them to classify document image objects automatically. This is done by developing classification algorithm that uses feature information. Neural Networks are powerful tools for handling problems of large dimension. The idea of combining extracted features and neural network is proposed. The interest in using artificial neural networks to classify document image has recently been confirmed by various works. To develop a classification algorithm using neural network, the data was be divided into a training set and a test set. This is done to use one set of gray documents to develop the classification scheme and a separate set to test the classification algorithm. After the data have been divided into the training and tested sets, work can begin on the development of the classification algorithm.

The proposed algorithms are described and their modules of application are demonstrated with several examples. The evaluation tests of these algorithms are given also.

2. The Proposed System
The proposed system consists of five modules : preprocessing, block erodes detection, block labeling, block features calculation, and neural network. This system is described in the following figure.

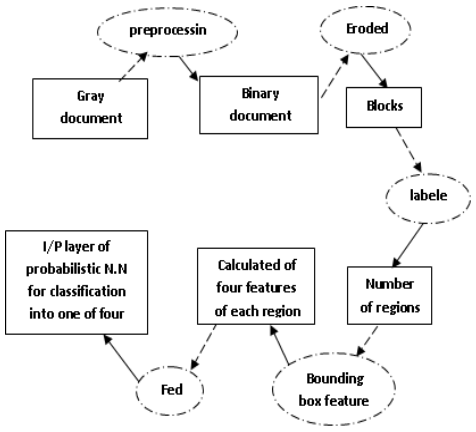


Figure (1) the proposed system

The process consists of two steps:

2.1 Analysis step: which contains the first four modules of the proposed system, these modules briefly described in the following paragraphs:

2.1.1Preprocessing
The input to this process is the gray image of document as the scanner captures it, then threshold is applied to the grayscale image to get the binary image which contain 0 and 1 values only. The output of this stage is a black and white document image with the same size as the original document.

2.1.2 The block eroding detection:

This module erode the binary document image which result from preprocessing module into blocks using the erode process.

2.1.3 The block labeling:

This module labels eroded blocks using the component labeling procedure. After being labeled, a block can be distinguished from another one based on its label, and each labeled block is called object.

2.1.4 The block features calculation:

In this step, the features of each labeled block (object) in a document image are calculated based on the bounding box. In this work, four features of each object is selected based on some observations with respect to size. total number of black pixels, total number of white – black or black – white of a block these features are

-Aspect ratio.
- Eccentricity.
-Extent.
-Bounding box_hight.

2.2 Documents classification Using Probabilistic Neural Network

The proposed document image classification algorithm can be performed by applying probabilistic neural network. This method needs the data base phase and other two phases for classification a document image, namely the training phase and the classification phase.

2.2.1 The Training Phase

Training phase deals with the problem of how to create reference of an document. The identification system needs to use knowledge about the given class of document. In most existing identification systems, this knowledge is used to perform good locating of an document to a given class. For every Identification system there must be a reference of documents concerned with in this system, it differs from one application to another. So, this reference will consist of information about the documents images, this information will be the class label for the document image.

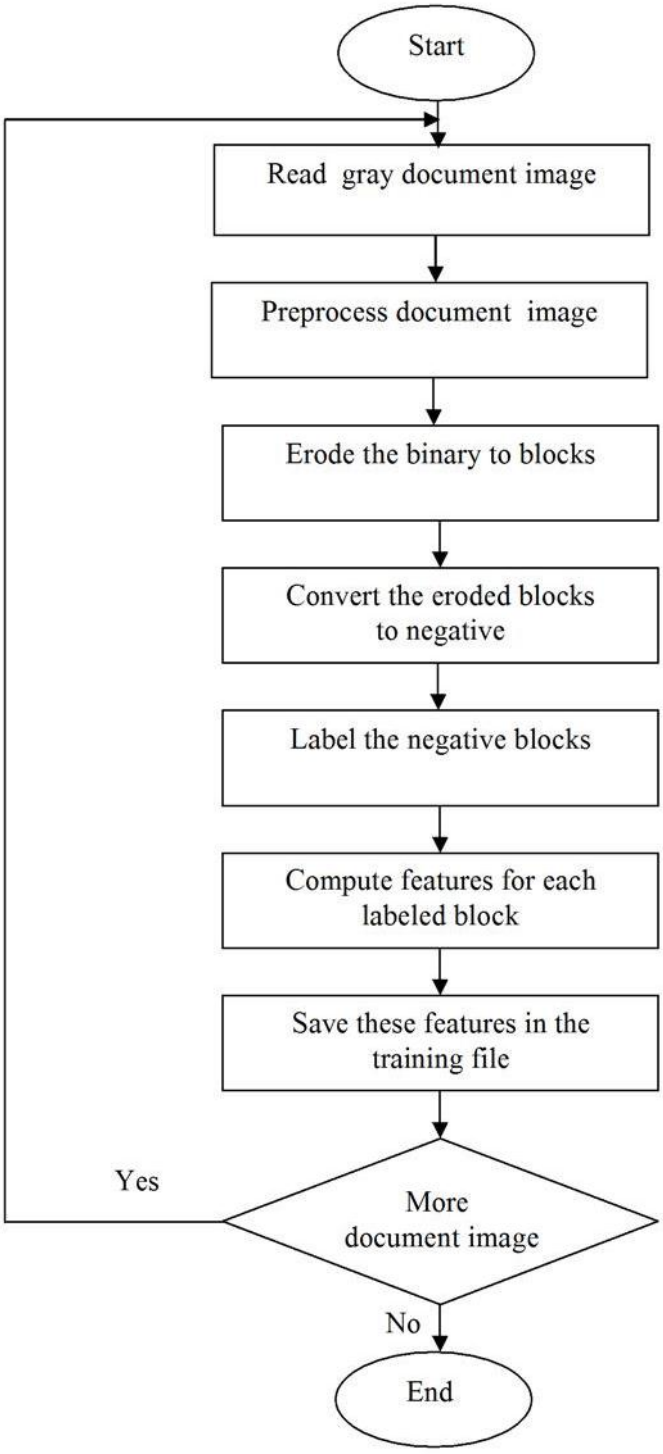
The Features_database that is used in generating the training information consist of documents images concerned within the application. The documents used in this approach are grayscale documents with any size. The document image is now in binary (black and white) form. The aim in this part is to decompose the document image into homogeneous regions. Homogenous in the sense that the regions have an attribute which satisfy a predefine criteria. The regions are text, picture, line drawing, map.

A statistics is calculated for a set of documents to obtain those features for each object

The erode (erosion process) is performed on the binary document image using (1 * 100) structural element connection. Two main objectives are realized in this process:

1. Reduction of the total connected components the process has to deal with and hence speeding the process.
2. The picture and map regions becomes dense and hence can be easily identified later.

see the figure (2) it is the block diagram of training phase.



Figure(2) the block diagram of The analysis phase

the following the training data and these features:

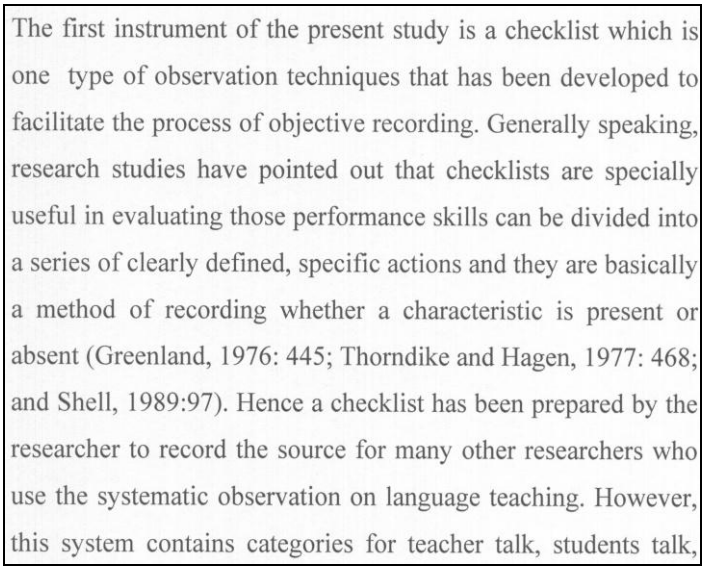

The first instrument of the present study is a checklist which is one type of observation techniques that has been developed to facilitate the process of objective recording. Generally speaking, research studies have pointed out that checklists are specially useful in evaluating those performance skills can be divided into a series of clearly defined, specific actions and they are basically a method of recording whether a characteristic is present or absent (Greenland, 1976: 445; Thorndike and Hagen, 1977: 468; and Shell, 1989:97). Hence a checklist has been prepared by the researcher to record the source for many other researchers who use the systematic observation on language teaching. However, this system contains categories for teacher talk, students talk,

Figure (3) text regions

Table 1 : Training Data of Text

No. of text	Aspect ratio	Extent	Eccentricity	Height of each block
1.	0.034175	0.15639	0.99956	69
2.	0.030708	0.16628	0.99965	62
3.	0.034192	0.15508	0.99959	69
4.	0.094656	0.12307	0.99567	62
5.	0.033218	0.15503	0.99957	67
6.	0.03173	0.16233	0.99964	64
7.	0.034209	0.15331	0.99962	69
8.	0.031111	0.15411	0.99963	63
9.	0.03373	0.15958	0.99959	68
10.	0.073171	0.13062	0.9974	63
11.	0.033267	0.15993	0.99959	67
12.	0.033764	0.15448	0.99957	68

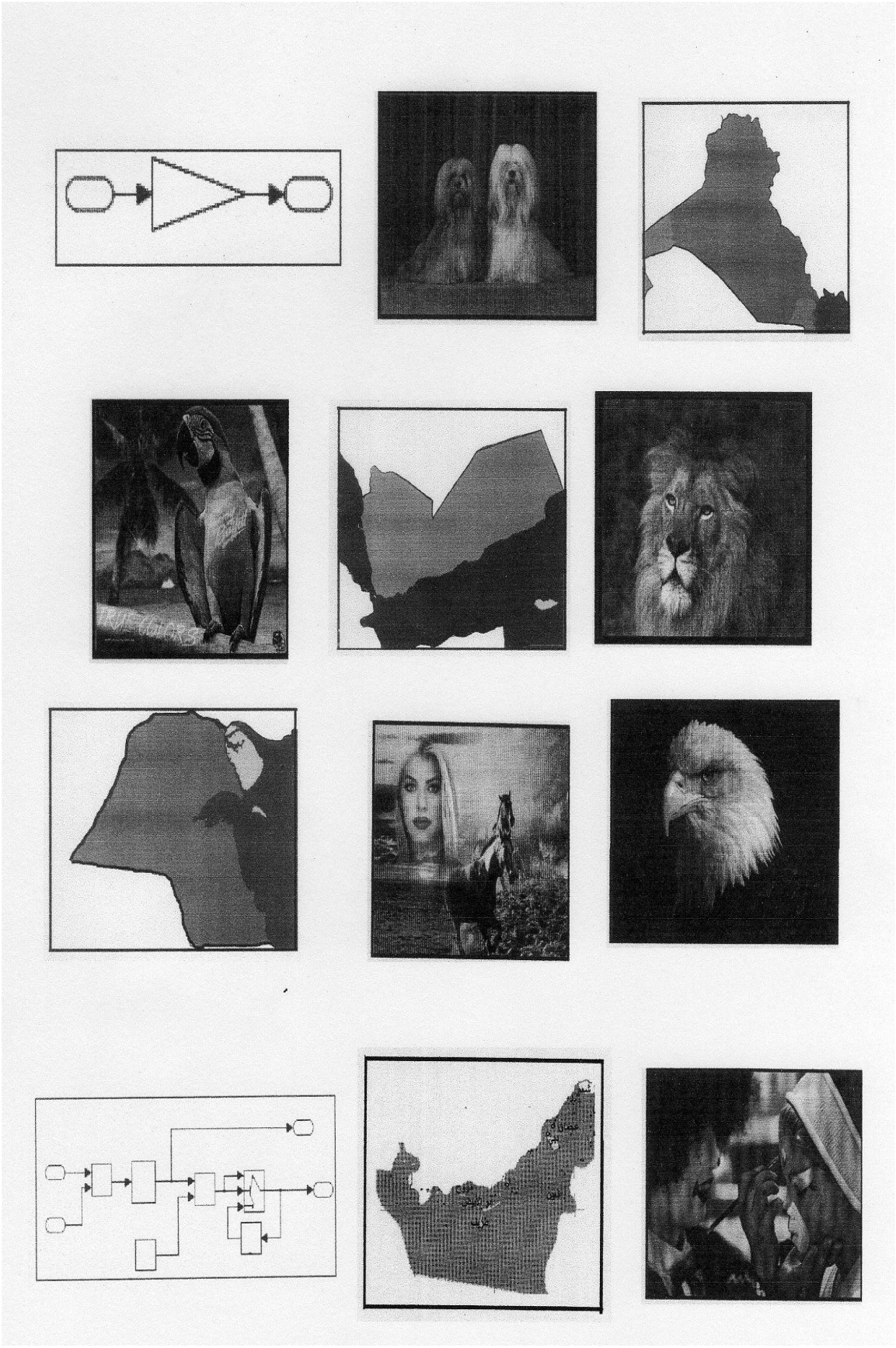


Figure (4) non -text regions

Table 2 : Training Data of non -text regions

No. of picture	Aspect ratio	Extent	Eccentricity	Height of each block
1.	0.71782	0.9658	0.69623	725
2.	0.75444	0.92839	0.65799	679
3.	0.67553	0.94436	0.73931	635
4.	0.66691	0.9429	0.74744	907
5.	0.66952	0.96829	0.74389	703
6.	0.66195	0.92832	0.75075	748
7.	0.65639	0.96512	0.7577	873
8.	0.67667	0.97674	0.74115	812
9.	0.63333	0.96478	0.77453	627
10.	0.75098	0.86077	0.66826	766
11.	0.66495	0.9784	0.74689	645
12.	0.67143	0.94255	0.74831	611

2.2.2 The Classification Phase

The first step in the classification phase is extracting the features of the gray document. These features are fed to the PNN to perform the classification.

2.2.3 The Classification Phase Via Probabilistic N.N. (PNN) Classifier

As mentioned before, the probabilistic N.N (PNN) is of the type Supervised, Feed forward, used for classification. It consists of four layers:

1- Input Layer: Consists of (4) nodes (length of each input vector

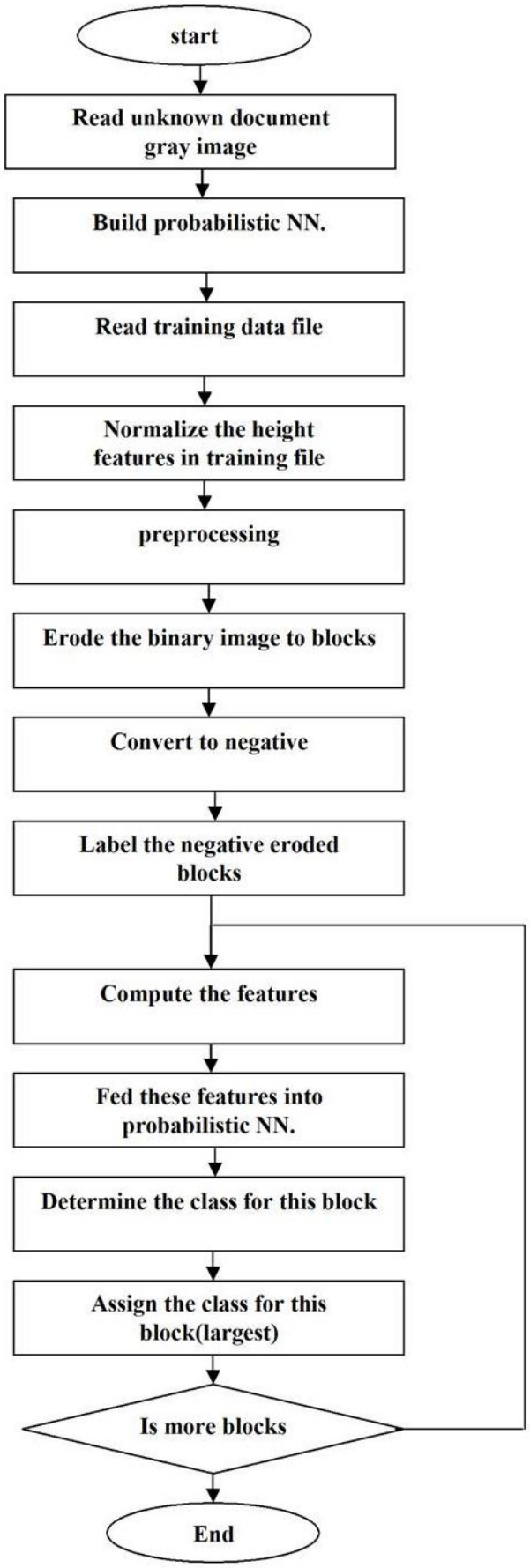
2 - Pattern Layer: Consists of (24) hidden nodes (number of training vectors).

There is one Pattern node for each training example. Each pattern node forms a product of the weight vector and the given example for classification, each neuron in the pattern layer computes a distance measure between the unknown input and the training case represented by neuron. where the weights entering a node are from a particular example.

3- Summation Layer: Consists of 2 hidden nodes (text and non_text). each summation node receives the outputs from pattern nodes associated with a given class, i.e. there is one neuron for each class, these neurons sum the values of the pattern layer neurons corresponding to that class in order to obtain and estimate probability density function of that class.

4 - Output Layer: Consists of 1 nodes (classes largest).

To make classification of unknown document image, the following steps will represent the proposed classification algorithm. The block diagram of classification phase will be as follows:



Figure(5) the block diagram for The classification phase

3. Experiment results:

3.1 Experiment 1 (image10006.bmp):

The document under consideration contains one text region, and four picture regions. In spite of this the algorithms have classified the document image regions correctly, as shown in the classified image layout in figure (6). The summary shows that there are (12) text lines in the document.

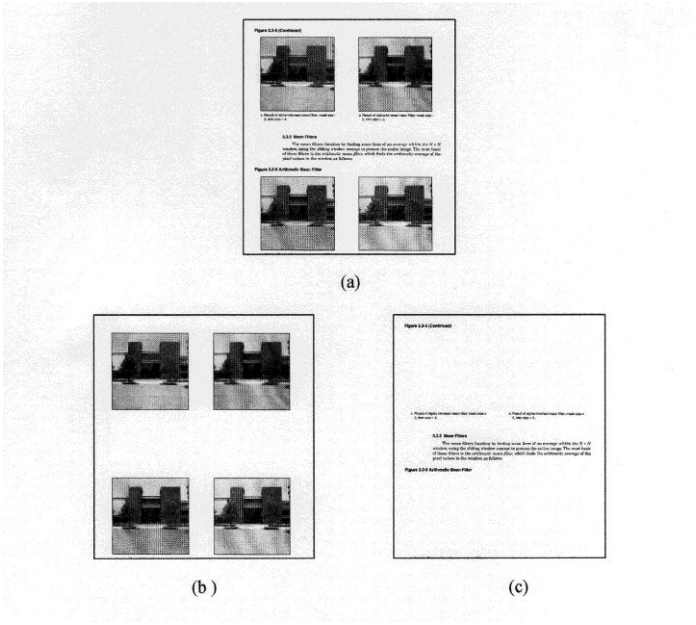


Figure (6) document image classification (a) original document image,(b) text region, (c) non -textual regions

3.2 Experiment 2 [upside down](image 10004 rotate. bmp): this document has been scanned upside down (rotate 180°).It contains one text region, and two picture regions. In spite of this the algorithms have classified the document regions correctly, as shown in the classified image layout in figure (7). The summary shows that there are (11) text lines in the document.

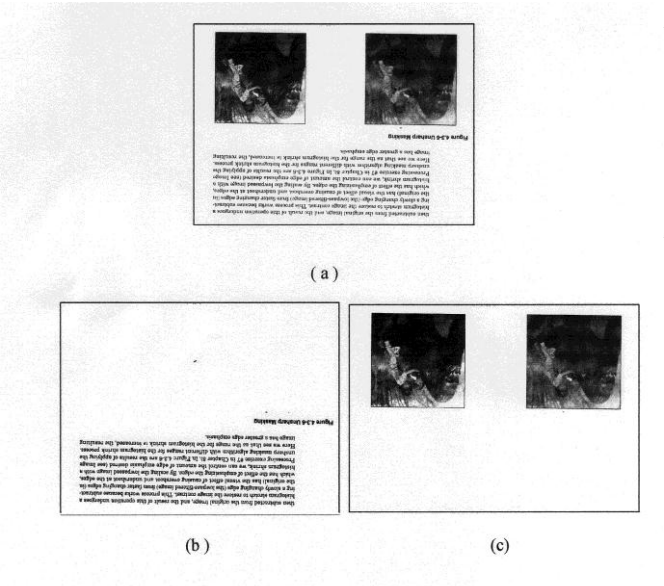


Figure (7) document image classification (a)original document image,(b) text region,(c) non -textual region

3.3 Experiment 3 (image10003.bmp):

The document under consideration contains text region, and one picture region. In spite of this the algorithms have classified the document regions correctly, as shown in the classified image layout in figure (8).

The summary shows that there are (17) text lines in the document

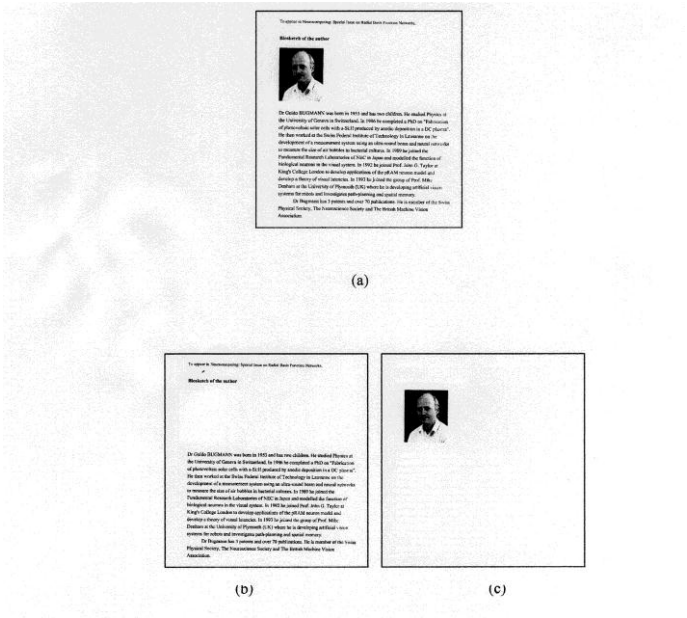


Figure (8) document image classification, (a) original document image,(b) text region,(c) non - textual region

4. Conclusion and Future work

There are several conclusion raised in the practical side of the research. These comments are discussed below:

1. Using probabilistic neural network gives the system more power because it needs short time and relatively little training set.
2. This algorithm classify the documents even if we scanned the document up side down.

5. Recommendation for Future Work

The following can be recommended for future work:

1. Implementing the document segmentation process by using the colored documents instead of the gray or the binary documents.
2. The document segmentation process can be implemented by using other type of neural net work such as Radial bases function, Back propagation, kohenen instead of the probabilistic used in this research

References

1. Umbaugh, Scott E "Computer Vision and Image Processing: a Practical Approach Using CVIP Tools" Prentice. Hall, Inc. 1998.
2. Wazir, Venus "An Investigation Into the Use of Neural Networks in Texture Classification" Ph. D. Thesis, Al-Nahreen University, 1999.
3. Gonzalez, Rafael C.& E., Richard"Digital Image Processing" Addison- Wesley Publishing Company, 2000.
4. Gloster, Clay & A.Macro, "Implementation of a probabilistic Neural Networks for Multispectral Image Classification on an FPGA Based Custom Computing Machine", 2001.
5. Le D.X, Thoma G, Weschler H, " Classification of Binary Document Images into Texture or Non-textual Data Blocks Using Neural Network Models", 2011.
6. Marinai Simone, Gori Marco, Soda Giovanni " Artificial Neural Networks for Document Analysis and Recognition",

تصنيف الوثائق المصورة الثنائية باستخدام الشبكة العصبية الاحتمالية

إيمان هشام عبدالرحمن

كلية التربية للبنات ، جامعة الموصل ، الموصل ، العراق

(تاريخ الاستلام: 17 / 1 / 2010 ---- **تاريخ القبول:** 2 / 4 / 2012)

الملخص

إن التكلفة المتناقصة لمكونات الحاسبة المادية ستمكننا من خزن ومعالجة الوثائق بشكل الكتروني, اليوم معظم الوثائق تخزن وتعالج وتعرض على الأوراق والتي هي أساس الكتب والمعاملات والصحف والمجلات.

و من اجل خزن وفهرسة ومعالجة مجموعة كبيرة من صور الوثائق يتطلب ذلك إجراء مجموعة من خطوات المعالجة. في هذا العمل استخدم نظام مقترح لتجزئة و تصنيف صورة الوثيقة الرمادية اللون الى منطقتين جزء نصي وغير نصي بالاعتماد على بيانات المقاطع التي تحويها. استخدم نموذج الشبكة العصبية الاحتمالية لهذا الغرض.

أولاً المعالجة المسبقة استخدمت لتحويل صورة الوثيقة الرمادية اللون إلى صورة الوثيقة ثنائية اللون (الأبيض والأسود) ثم استخدام عملية التآكل لتحويل الوثيقة ثنائية اللون إلى وحدات، الوحدات الناتجة تقسم الى عدد من المناطق باستخدام البرنامج الفرعي للعنونة.

أربع صفات لكل منطقة تحسب بالاعتماد على المستطيل الذي يحوي تلك المنطقة بعدها يتم إدخال الصفات المحسوبة الى وحدة الإدخال في الشبكة العصبية الاحتمالية لغرض تصنيفها إلى منطقتين اما نصية او غير نصية.

بعض صور وثائق الرماديةِ اللون تم استعمالها لكي تَختبرَ النظام المُقتَرَح. أخيرا عشرة تجارب نفذت وصنفتها الخوارزمية بشكل صحيح

الكلمات الدالة: تصنيف الصور الثنائية، الدالة العصبية الاحتمالية.