تصنيف صورة الوثيقه بواسطة الشبكة الاحتماليه العصبيه

م .م. وداد عبد الخضر ناصر
الجامعة المستنصريه
كلية التربية / قسم الحاسبات

م أمل عباس كاظم
الجامعه المستنصريه
كلية التربية / قسم الحاسبات

المستخلص:

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.

ومن اجل خزن وفهرسة ومعالجة مجموعة كبيرة من صور الوثائق يتطلب ذلك إجراء مجموعة من خطوات المعالجة. في هذا العمل تم استخدام نظام مقترح لتجزئة و تصنيف صورة الوثيقة الرمادية اللون الى مناطق 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.

تم حساب اربع صفات لكل منطقة بالاعتماد على الصندوق الذي يحوي تلك المنطقة بعدها يتم إدخال الصفات المحسوبة الى وحدة الإدخال في الشبكة العصبية الاحتمالية لغرض تصنيفها إلى واحدة من منطقتين وهي نص و صورة .

تم استعمال بعض صور وثائق الرماديةِ اللون لكي تَختبرَ النظام المُقتَرَحَ. الاختبارات نجحت في تصنيف جميع وثائق الاختبار .

Document Image Classification Using Probabilistic Neural Network

Abstract:

The decreasing cost of hardware will eventually enable commonplace the storage and process of documents by electronic means. However, today most documents are begin saved, processed, and presented on papers. Paper is the primary medium for books, journals and newspapers. In order to properly archive, index and process a large number of document images, several challenging processing steps must be completed. In this work, a proposed system is used to segment and classify the gray document image to regions 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 (text, picture). Some gray documents images are used in order to test the proposed system. The tests have successfully classified all test documents.

1- Introduction

Computer imaging can be defined as the acquisition and processing of visual information by computer.

we can separate the field of computer imaging into two primary categories:

1- Computer vision.

2- Image processing.

In computer vision applications, the processed (output) images are for use by a computer, whereas in image processing applications, the output images are for human consumption. The major topic within the field of image processing includes image restoration, image enhancement, and image compression. Image analysis is often used as preliminary work in the development of image processing algorithms [1,2].

Image analysis is a process of discovering, identifying, and understanding patterns that are relevant to the performance of an image – based task[3].

Documents classification is very important and can be used in many applications. The aim of this research is to propose algorithms to classify the gray document image into text, picture [4]. 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 interest in using artificial neural networks to classify document image has recently been confirmed by various works [5,6].

2 . System Overview

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

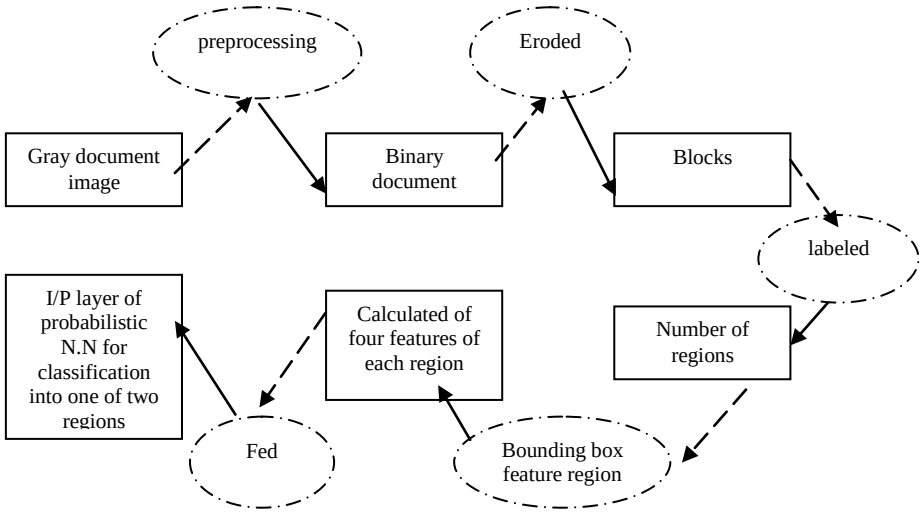


Figure (1): The Proposed System

The five modules of the proposed system are briefly described in the following paragraphs:

2.1 Preprocessing

The input to this process is the gray image of document as the scanner captures it, the output of this stage is a black and white document image with the same size as the original document. Applying a threshold to grayscale image to get black and white i.e. the image matrix now contains only (0) and (1) values. Thresholding in its simplest form means to classify the pixels of a given image into two regions (e.g. objects and background) [4,7].

2.2 The Block Eroding Detection :

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

Erode process performs erosion on binary image . It automatically takes advantage of the decomposition of a structuring element object [4,7] .

2.3 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.

2.4 The Block Features Calculation:

This step, 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 are 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.

3. Documents Classification Using Probabilistic Neural Network

The proposed document image classification algorithm can be performed by applying probabilistic neural network. This method needs three phases: the data base phase ,the training phase , and the classification Phase.

3.1 Database phase

The database that is collected for gray document classification which consists of many gray documents images, represents four classes (text, picture, line drawing, map). These documents images represent a different kinds of documents. These documents are converted by scanner device to computer images.

3.2 Training Phase

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

The Features database that is used in generating the training information consists of document images concerned within the application. System documents are grayscale documents with any size. When the document image is chosen for training file, it needs two work phases : (Preprocessing and Region identification)

The document image is now in its 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 a attributes which satisfy a predefine criteria. The regions are text, picture, line drawing and map.

A statistics are 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.**

- **<u>Proposed Algorithm For Training Phase</u> :**

Step(1) Read an input document, which is named
Step(2) Convert the gray document into binary document
Step(3) Erode the binary document into eroded blocks
Step(4) Convert the eroded blocks into negative
Step(5) Label the negative eroded blocks
Step(6) Compute features for each negative labeled block
Step(7) Save the features in a training file
Step(8) If there are more documents image, go to step (1)
Step(9) Close database file
Step(10) Return

- **The Flowchart of Training Phase Algorithm :-**

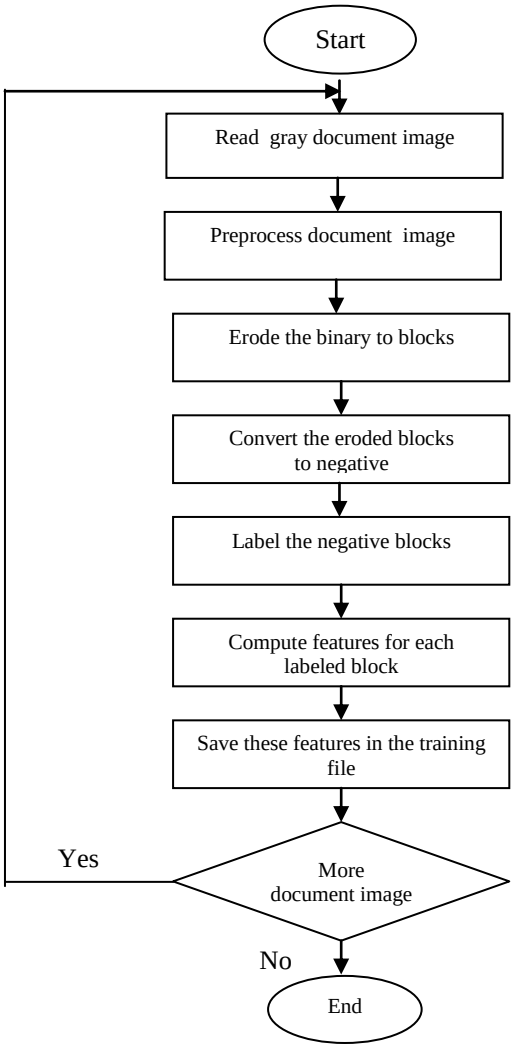


Figure (2) : Training Phase Flowchart

The Training Data and their 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,

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

Table 1 : Training Data of Text

Figure (3): Picture Regions

Table 2 : Training Data For Pictures of Figure (3)

No. of picture	Aspect ratio	Extent	Eccentricity	Height of each block
13.	0.71782	0.9658	0.69623	725
14.	0.75444	0.92839	0.65799	679
15.	0.67553	0.94436	0.73931	635
16.	0.66691	0.9429	0.74744	907
17.	0.66952	0.96829	0.74389	703
18.	0.66195	0.92832	0.75075	748
19.	0.65639	0.96512	0.7577	873
20.	0.67667	0.97674	0.74115	812
21.	0.63333	0.96478	0.77453	627
22.	0.75098	0.86077	0.66826	766
23.	0.66495	0.9784	0.74689	645
24.	0.67143	0.94255	0.74831	611

3. 3 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 classification.

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 (48) 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 4 hidden nodes (number of classes). Each summation node receives the output 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.

proposed classification algorithm:- The

Step(1) Read an input image.

Step(2) Build the probabilistic Neural Network.

Step(3) Read the training data to train the neural network.

Step(4) Normalize all the heights of bounding boxes (where height is the fourth feature that can be extracted from the gray document in the training file).

Step(5) preprocessing

Step(6) Erode the binary document (using 1*100 structural element) into eroded blocks.

Step(7) Convert the eroded blocks into negative eroded blocks.

Step(8) Label the negative document.

Step(9) Compute the features for each labeled block
Step(10) Input the features' of new block which results from step (9) to the neural network.
Step(11) Find the estimated PDF for each hidden node in pattern layer.
Step(12) Find the sum of each node in summation layer (sum of estimated PDF for each class).
Step(13)Find the probability of each class by dividing the sum of estimated PDF for each class over the sum of all estimated PDF.
Step(14)Assign the unknown document block to a class that is fried from PNN (with largest probability).
Step(15) If there is more blocks? Go to (Step9).
Step(16) End.

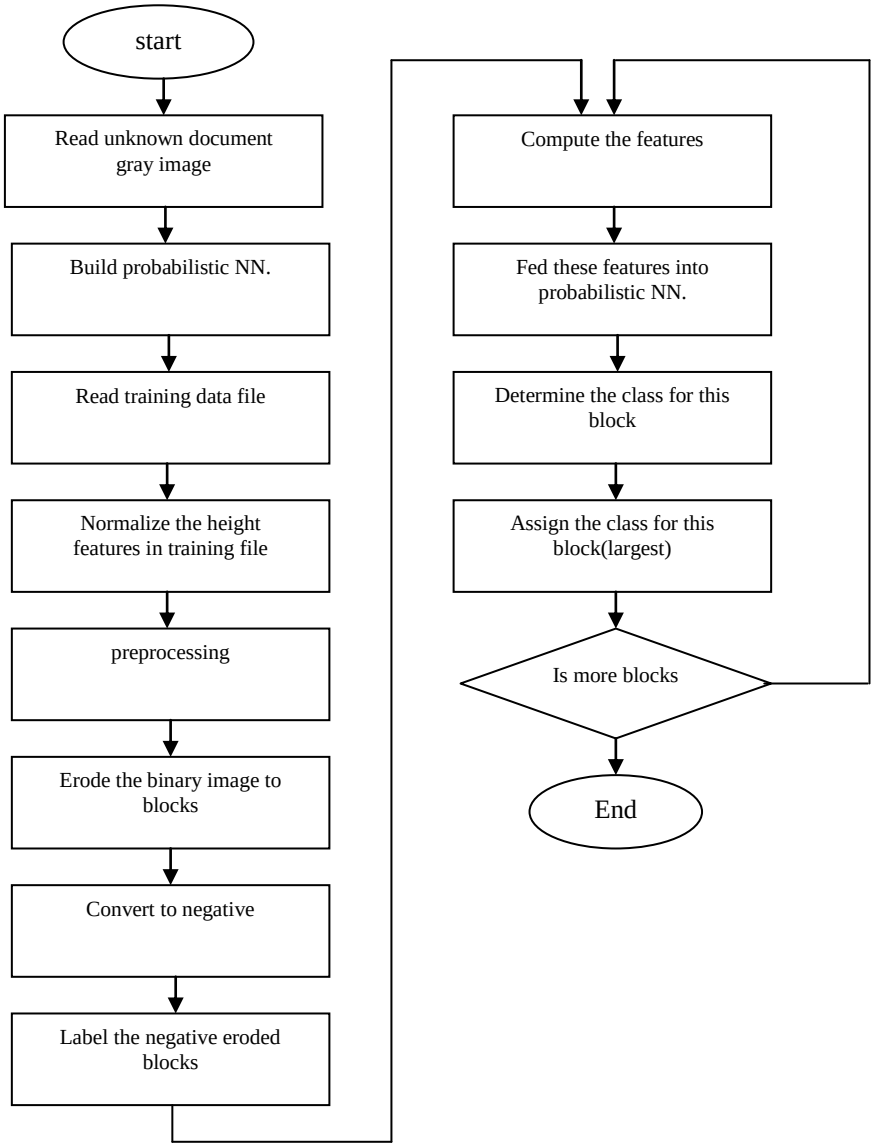
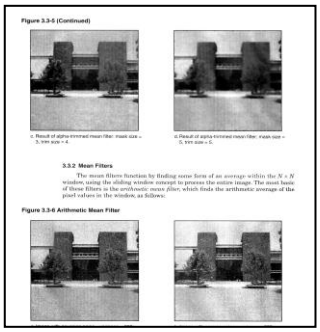


Figure (4): Classification Flowchart

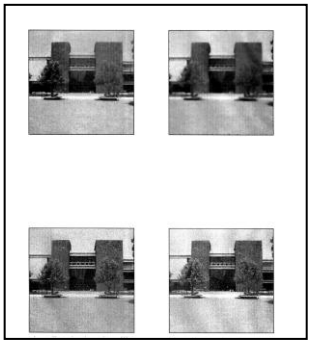
4-Results

4.1 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 (5). The summary shows that there are (12) text lines in the document.



(a)



(b)

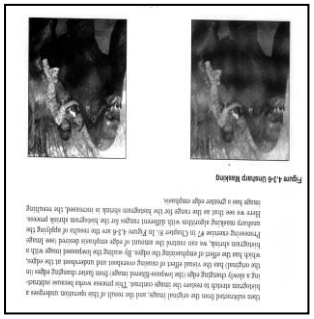
(c)

Figure (5): shows document image classification
(a) original document image, (b) picture region,(c)text region

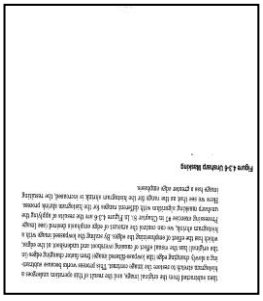
4.2 Experiment 2 [upside down](image10004rotate.bmp):

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 is shown in the classified image layout in figure (6).
The summary shows that there are (11) text lines in the document.



(a)

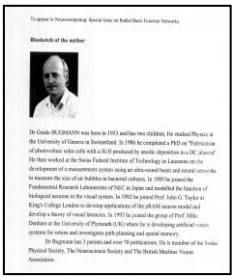


(b) (c)

Figure (6) shows document image classification
(a)original document image,(b) text region,(c)picture region

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



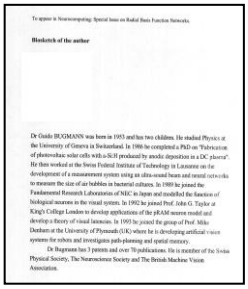
To appear in Neurocomputing: Special Issue on Radial Basis Function Networks.

Blokstch of the author

Dr Guido BUGMANN was born in 1953 and has two children. He studied Physics at the University of Geneva in Switzerland. In 1986 he completed a PhD on "Fabrication of photovoltaic solar cells with a-Si:H produced by anodic deposition in a DC plasma". He then worked at the Swiss Federal Institute of Technology in Lausanne on the development of a measurement system using an ultra-sound beam and neural networks to measure the size of air bubbles in bacterial cultures. In 1989 he joined the Fundamental Research Laboratories of NEC in Japan and modified the function of biological neurons in the visual system. In 1992 he joined Prof. John G. Taylor at King's College London to develop applications of the pRAM neuron model and develop a theory of visual latencies. In 1995 he joined the group of Prof. Mike Denham at the University of Plymouth (UK) where he is developing artificial vision systems for robots and investigates path-planning and spatial memory.

Dr Bugmann has 2 patents and over 70 publications. He is member of the Swiss Physical Society, The Neuroscience Society and The British Machine Vision Association.

(a)



To appear in Neurocomputing: Special Issue on Radial Basis Function Networks.

Blokstch of the author

Dr Guido BUGMANN was born in 1953 and has two children. He studied Physics at the University of Geneva in Switzerland. In 1986 he completed a PhD on "Fabrication of photovoltaic solar cells with a-Si:H produced by anodic deposition in a DC plasma". He then worked at the Swiss Federal Institute of Technology in Lausanne on the development of a measurement system using an ultra-sound beam and neural networks to measure the size of air bubbles in bacterial cultures. In 1989 he joined the Fundamental Research Laboratories of NEC in Japan and modified the function of biological neurons in the visual system. In 1992 he joined Prof. John G. Taylor at King's College London to develop applications of the pRAM neuron model and develop a theory of visual latencies. In 1995 he joined the group of Prof. Mike Denham at the University of Plymouth (UK) where he is developing artificial vision systems for robots and investigates path-planning and spatial memory.

Dr Bugmann has 2 patents and over 70 publications. He is member of the Swiss Physical Society, The Neuroscience Society and The British Machine Vision Association.

(b)

(c)

Figure (7) shows document image classification, (a)original document image,(b) text region,(c) picture region

5- Conclusions :

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 upside down.**

References:

1. **Umbaugh, Scott E. "Computer Vision and Image Processing: a Practical Approach Using CVIP Tools" Prentice. Hall, Inc. 1998.**
2. **Starck, J. L. & Murtagh, F. & Bijaoui, A. "Image Processing and Data Analysis " The Multiscale Approach Cambridge University Press, 1998.**

3. **Wahl, F.M., Wong, K.Y. and Casey, R.G. "Document Analysis System" IBM Journal of Research and Development, 1982.**
4. **Dave A. D. , Tompkins and Faouzi Kossentini, "A Fast Segmentation Algorithm for Bi-Level Image Compression using JBIG2". Internet Report, 1998. http : // spmg.ece.ubc.ca**
5. **Le , D. X., Thoma, G., Weschler H. , "Classification of Binary Document Image into Texture or Non-textual Data Blocks Using Neural Network Models", Internet Report 2002.http://archive.nlm.nih.gov/pubs/doc_class/mv.php Updated February 13, 2002.**
6. **Al-Adhadh, F.A. "Computer Vision for Machine and Hand Written Documents" Ph.D.Thesis, The National Computer Center, Baghdad,2003.**
7. **Gonzalez, Rafael C.& E., Richard . "Digital Image Processing" Addison- Wesley Publishing Company, 2000.**