Efficient Text Segmentation System for Heterogeneous and Complex Documents

A Dissertation

Submitted

By

Lalit Kumar

11311350

To

Department of Computer Science & Engineering

In Partial Fulfillment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science Engineering

Under the Guidance of

Mr. Harsh Bansal

Astt. Professor

(May 2015)

LOVELY PROFESSIONAL UNIVERSITY

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student	: LALIT KUMAR	Registration No	: 11311350
Batch	: 2013-2015	Roll No	: RK2307A69
Session	: 2014-2015	Parent Section	: K2307

Details of Supervisor:

Name	: HARSH BANSAL	Designation	: Assistant Professor
UID	: 16866	Qualification	: M.TECH
		Research Exp.	: 2 years

Specialization Area: Database (pick from list of provided specialization areas by DAA)

Proposed Topics:-

1. EFFICIENT TEXT SEGMENTATION SYSTEM FOR HETEROGENEOUS AND COMPLEX DOCUMENTS
2. AUTO COLUMN SEGMENTATION IN HADWRITTEN DOCUMENT.
3. NOISE REMOVAL APPROACH FOR IMAGE PROCESSING.

Harsh 16866

Signature of supervisor

PAC Remarks:

first topic approved

11/29

APPROVAL OF PAC CHAIRMAN Signature: 2-9-14 Date:

*Supervision should finally encircle one topic out of three proposed topics and put up for an approval before Project Approval Committee (PAC).
*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.
*One copy to be submitted to supervisor.

ABSTRACT

In the document image analysis document segmentation is very important step. Document segmentation is the process in which we segment the document which contains the heterogeneous data means data like printed text, handwritten text, graph etc. We do the document segmentation because our optical character recognition system is unable to recognize the whole document with multiple data type so before the recognition we have to apply the document segmentation so to define the each region correctly. We would be using document segmentation on the handwritten bills which contain the heterogeneous content thereby segmenting the text and non- text region and the text into printed text and handwritten text and then we classify the text region into printed text and handwritten text and map it to the text document.

Keywords: Segmentation, Recognition, Handwritten Bills

ACKNOWLEDGEMENT

First of all, I would like to thank my Almighty God, who has always blessed me and for giving me strength to do this work.

I wish to express my deep gratitude to my guide, Mr. Harsh Bansal, for his generous guidance. His guidance and support throughout all stages of the thesis process enabled me to conduct the research. Without his support, I would not be possible to complete this program.

Name : Lalit kumar

Registration No. 11311350

DECLARATION

I hereby declare that the dissertation entitled "**Efficient text segmentation system for heterogeneous and complex documents"** submission for the M.Tech degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Investigator

Reg. No. 11311350

CERTIFICATE

This is to certify that Lalit Kumar has prorating M.Tech dissertation proposal "**Efficient text segmentation system for heterogeneous and complex documents"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date :

Signature of advisor

Name: Harsh Bansal

Assistant Professor

UID: 16866

Lovely Professional University

TABLE OF CONTENT

LIST OF FIGURES

Chapter 1
Introduction

Natural language processing is used to build the system which can understand and generate the natural language. The basic idea is to build these kinds of systems on the bases how human understands and communicates. We make the many system on the behalf of this idea like machine translation system, man machine interface, speech understanding, text analysis system which can understand the printed as well as handwritten text.

1.1 Application of Natural Language Processing

II. **Text based application:** Involve written text processing like from the newspaper, email, report there is some benefit of this approach like

- Find the appropriate document from the large amount data like to the certain type of book from the library.
- Extract the appropriate information from the certain data like extract the all stock transaction for the particular day.
- Translate the one document language to another language.
- To summarizing the large amount of data like to summarize the 1000 page document to 5 page document by choosing the appropriate key word from the documents .

III. **Dialogue based approach:** In this human communicate with machine by spoken language but sometime it also done with keyboard typing. There is some application of this approach like

The first application is query to the database in the form of natural language.

- The feedback is also in the form of natural language like we ask the library management system in natural language for the particular author present in the library or not and the answer is also in the form of natural language.

- Tutoring system where systems teach the student and student can also interact with the system in the form of natural language.

But there is some problems in the dialogue application like understand the language because one language spoken by the different man is different so to understand that language we have to recognize and classify the dialogue properly.

1.2 Segmentation

Segmentation is the process which done before the recognition process because it change the object into digital format and then divide it into the proper segment so that it is easy for the recognition to this object.

In the natural language processing we divide the segmentation into basically three parts

1.2.1 Image segmentation

Image segmentation is to analysis the image and then partition the image into homogenous r in homogenous region. image segmentation is very important part of the image analysis because show the important and the interesting area of the image for example to segment the MRI image which contain the brain tumor with the help of the image segmentation we can the highlight the brain tumor part for which we are interesting.

1.2.2 Text segmentation

Text segmentation is the preprocessing step for the optical character recognition. Text segment is the process in which partition the text into the particular reason to easily recognizable. In the text segmentation we include the printed as well as handwritten text.

1.2.3 Document segmentation

Mostly of our document contain the heterogeneous data means in the text, image, and graphs together so for to recognition to this kind of table we need to segment the document to the particular region textual data to the textual reason and image to their image region so it is easy to recognition the document.

1.3 Document image analysis

Document image analysis is the process to analyze the document by the computer as the human can do. The main of this process is that computer can understand the document page as it understands the other media and made the office to ‘paperless office’.

The main objective of document image analysis to recognize the textual as well as graphical parts in the document as the human can understand and extract the useful knowledge from the document.

Document mainly containing two types of data. One in the form of textual and other is the graphical form.

I. **Textual processing**: - Textual processing deal with textual part of the document which contain the words, text lines, skew angles of the different text and the other kind of stuff like to differentiate between the printed text and the handwritten text. The work of textual processing not to only deal with the textual part but also to recognize this textual part with the help of OCR.

II. **Graphical processing**:-Graphical processing deal with the graphical component of the document which contain the different kind of graphical structure like tables, images, logos and straight lines between the text section. Most of these graphical components include lines so in the graphical processing it deals with the line detection, line fitting and line curve. The third component of the document is the picture but it is the part of image processing which is not the part of document image processing.

1.3.1 Hardware Advancement and Evaluation in DIA

Document image analysis is start through from the find out the new technology like digital signal processing and the document image processing. Digital signal processing firstly can be in the use with the development of fast computer and the fast processing algorithm like Fourier transform in 1960 and the main motive of this algorithm to process the one dimensional signal like voice. In 1970 with the development of large computer and more large memory for the storage of data

they gone to be performed the processing of two dimensional data for example digital picture.

In 1980 document image analysis technique grow rapidly this is all due to the hardware advancement with the appropriate hardware the same work going to be done on the low cost and less time. To understand this we are going to take the example that firstly to process the audio signal we need the frames of 256 samples and 512*512 digital image vision sizes. On the other hand to process the document images for example a business letter 2550*3300 pixel which means 300 per inch.

1.3.2 Common Sequence and Steps in DIA

The first step in DIA to capture the data in the form of document image and the next step is to change this image into the pixel form. Pixel is basic unit of an image and most of the time the pixel has two types of values 0&1 for the binary image and 0 to 255 for the grey scale image which also represent the color image. There we must have to understand that document image contain only the raw data from which we have to extract the useful information.

For example an on -off pixel show an image for computer it is an image or it can be string of data but for human it can be any character.

I. **Processing at pixel level:** At this level of stage all the operations like segmentation, binarization and noise reduction is been performed.
    - The first of all the operation which are going to performed first is the binarization in this operation we can change any gray scale or color image into the 0&1form means into the form of white and black. The basic objective of binarization is differentiating the foreground and background pixel of the image so that we can easily access the useful knowledge from it. For example to extract any text from any image we can perform the binarization operation with the help of this we can easily understand that which part of the image contain the text part.

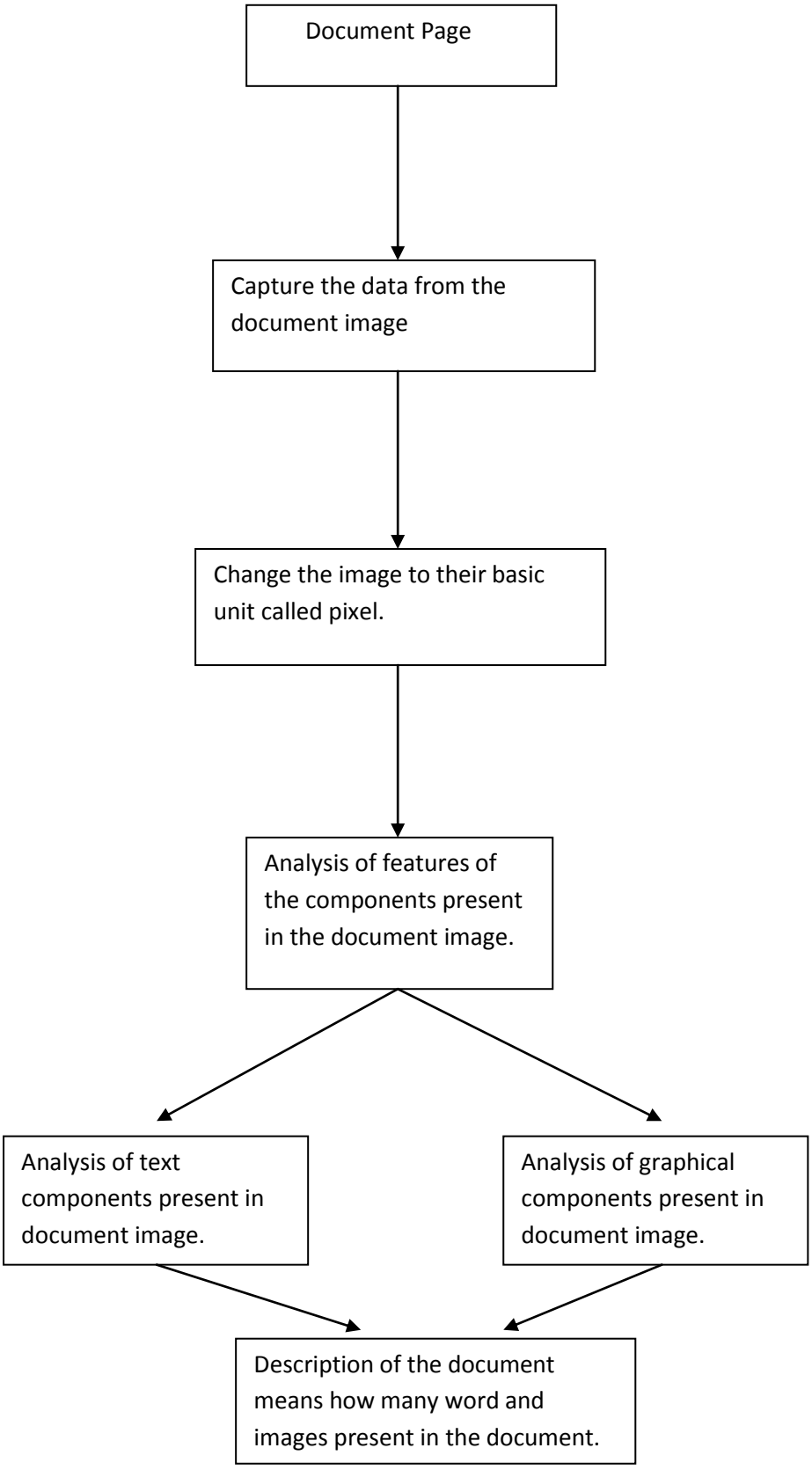


Figure 1.1 Sequence of steps in DIA

- Noise removal is the next which we are perform in the document image analysis. The basic reason of presence of noise in the document image due to the image transmission, photocopy mistakes and aging .the most common type of noise in document is the salt &pepper noise. The reason of this noise due to the dirt, speckle noise etc. this type of noise randomly distribute all over the document image means black specks on the white background and white specks on the black background and to remove this noise most commonly used filtered is the median filter this filter take the median of the background pixel and replace the most unmatched value with the median pixel value.
- Signal enhancement is the other kind of noise removal technique but it is basically used in the graphical component the basic function of this technique is to fill the lines which otherwise are suppose to be continues.
- Segmentation is the next operation which is taking place after the noise removal. In this step document are divided into the two parts one the parts contain the textual components and the other parts contain the graphical component. Textual component contain the data like words, caption and titles. On the other hand graphical component include the data like tables, lines, drawings etc. for example to segment any table which contain the text first of all we going to separate the text from to the table and then we are going to separate the template of the table which are belong to the graphical component.

II. **Feature level analysis:** In the feature analysis of text image global feature describe the characteristics of the each word, each font and the other feature like how many number of crossing in each word and this method is followed by the OCR system for recognize the text.

In the feature analysis of graphical image global feature describe the line width, minimum line length and on the other hand local feature describe the straight line, corners and position of the circles, rectangle etc which are present in the document image.

III. **Text and graph recognition:** In this part we going to differentiate between which part of the document is belonging to the graphical component and which part belong to the textual component. In this part we not just only differentiate

but also recognize the textual and graphical component. In the textual part we analysis the word, characters and printed and handwritten text and in the graphical part we analysis the graphical component like company logos, picture and tables etc.

1.3.3 BRIEF INTRODUCTION OF EACH STEP FOLLOWED IN DIA

I. **Document image preparing:** This is the first step document image analysis. In this step we are going to scan the document image and change into the 0&1 or 0 to 255 pixel value. Threshold is an also the important step in image processing the basic work of this step to minimize the gray scale value and remove the noise from the image.

- **Thresholding:** In the document image processing us change the image into binary image and the function of the binary is to differentiate the foreground and background pixel. Foreground pixel is called the area of interest but due to some error like bad photography, poor scanning of image some of the background pixel taken as foreground pixel so to remove to this kind of pixel we do the thresholding the basic function of this is to remove that pixel which has the threshold value above then this but to do this we have to clearly choose the threshold value.

- **Noise reduction:** After the binarization of the image our next step is to remove the noise from the document image. The common type of the noise in the document is salt& pepper noise. In this type of noise ON pixel is present in OFF pixel region and OFF pixel is present in the ON pixel region. To remove this noise we mostly commonly use the median filter and median average filter. The basic function of the noise reduction process to ease the process of recognition and also noise reduction process small the size of the image so less time talen to to process the image.

II. **Find appropriate feature:** After the pixel level processing the next part in document image analysis is to find appropriate feature from the image. For example to find the feature of line we have to find that this line is straight line or in the cursive form in the same way that the human can do. In this process we

find the feature of both the graphical as well as textual component. This process make easy to the OCR system for recognition. In the feature analysis we have the different method for analysis of the component like polygonalization this is the method in which we analysis the feature of lines. In this method we analyze the curve and mapped it with the original curve and differentiate that how that how that both curve relate to each other.

- **Recognition shape description:** In this step we are going to describe the shape of different object. In the shape description we are going to defined that which component are belong to textual component and which are belong to graphical component. We all know graphical component include the long lines, circles etc and on the other side textual component include the fonts. To differentiate between the two objects we use the methods likes curve detection and curve fitting. To differentiate between the longest font and smallest fonts we use the technique of connected components.

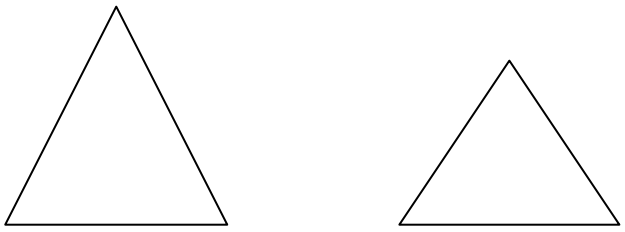


Fig 1.2 To differentiate the size of the object we check the number of on pixel, length of contour and area of bounding box.

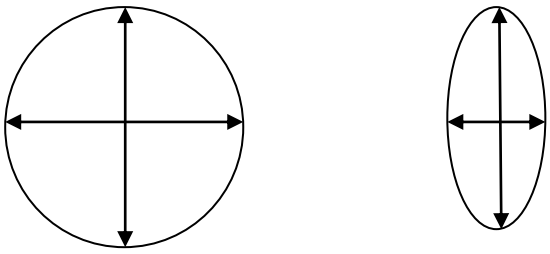


Figure1.3 To find the compactness between the two object the mostly common used method to find out the ratio between major and minor axes.

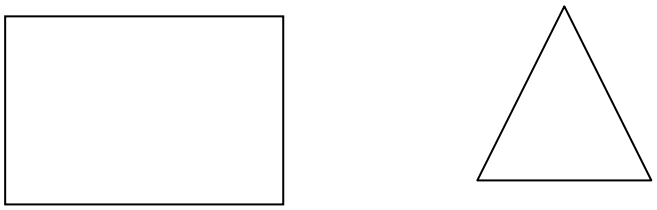


Figure1.4. To differentiate between the shapes we check the symmetry and asymmetry structure of the object.

III. **Map the text on excel sheet Recognize the component of text line:** This is applied to find out the textual component in the document. It also help to find out that which block contain the text textual component and on which position the textual component is present. Both OCR and format analysis is related to each other because on technique is used to find out the position of the textual component in the component and other is used to find out the meaning of that word but the technique complement to reach other. In this method we used both the technique.

- **Skew estimation:** In the skew estimation technique we are going to find the skew angle of text lines with respect to the page layout. The document has the zero skew value if horizontal and vertical text lines are parallel to the page layout. Skew estimation is done before to display the document onto the screen.
- **Layout analysis:** After the skew segmentation we perform the layout analysis. One of the layout types is the structured layout analysis. In this layout analysis we segment the document into different components like word, text line etc. we done the structured analysis by two methods one is the top down method and other is the bottom up method. In the top down method we are started from the largest component. In this we divide the document into column format and after column is divided into the paragraph and paragraph is divided into the text lines and text lines are divided into the word. On the other hand in bottom up we start from the lowest component of document like character and then character are merged into the words and words are merged to create the paragraph.
- **Printed character recognition:** This is the most basic task for OCR to recognize the printed task. In the early days OCR recognize this character by matching with the predefined template because each typewriter has the same character size and shape but after the innovation of new technology like laser printer and page typewriter the task of character recognition is more difficult because each character of this

typewriter are different to another printed character because each character has their different weight, different size and shape so this is difficult to map this characters with predefined template so to overcome this problem OCR system used the feature analysis method in this method we analysis the feature of the each character and train the system on that features.

- **Segmentation of the character:** This is preprocessing step before the character recognition. In this step we divide the character into their own region so that it is easy for OCR system to recognize it. As good as the segmentation so there is the more chance for good recognition. The basic character segmentation method is the vertical projection method in this method text paragraph is vertically divided into the words then into the character.
- **Noise reduction:** It is also helpful for the segmentation task because some time due to the reason of the noise character are wrongly segmented so it is very important to use the right noise reduction technique before to segment the character.
- **Handwritten character recognition:** Handwritten recognition is new area of research now these days because there is lots difference between the handwritten recognition and the printed text recognition. The reason behind this in printed character each character in the well structured form but on the other hand in handwritten character each character in cursive form and each user has their own way to write a character so it is difficult to recognize the same character into different ways. There is two type of handwritten character recognition one is off-line character recognition and other is the on-line character recognition. During the on-line recognition we also going to check the pin speed and tip size of the pin but in the off-line recognition we simply scan the document as we done in the printed text recognition.

IV. **Recognize the graphical component in document:** In this part we deal with the graphical components of the document. Graphical component like tables, drawings, graphs etc. The main objective to recognize the graphical component is that by doing this we can semantically describe the graphically image. This is

very important to know the graphic component semantically by doing this we can make change in the graphical images like we done in the CAD.

1.3.4 Other Application of DIA

Document image analysis used in very different area of application. There are some of the applications like:-

- **Analysis of fingerprints:** Analysis of fingerprints is the one of the application of the recognition of the graphical component. First of all in fingerprint processing we change the image into grey scale and then done the pixel level processing on the image for removal of noise from image. After the removal of the noise we change it into the binary image and detect the microscopic pattern and form a direction map image and map it with the patter
- **Printed music processing:** To done the printed music processing we use the top down processing method in this method we extract the musical notation from the complex document. To done the proper extraction we have to do the proper segmentation.
- **Extraction of information from microfilm images of punching cards:** There is lots of old data present on the microfilm images of punching cards but this data is useless until it can be change into the computer readable form. But with the help document image analysis we can do this.

  Document image analysis contains the semantic analysis step which can help to change the data present on the original punching card to the computer readable form.

1.4 Document Page Segmentation

Document page segmentation is the most important step in the document image analysis to understand the significance of the document page segmentation in document image analysis we understand the concept of document image analysis.

Now these days lots of data present on the document so the work of document image analysis to analysis the document which contain the textual and graphical data and extract the useful from it as the human does .there is some further part in document analysis like

- Document processing in this step we scan the whole document is to be scanned and with the help of document segmentation we can divide it into further textual part and graphical part.
- Textual processing in this only the textual data is to be processed which contain printed as well as handwritten text. Further the textual processing is divide into two parts OCR and page layout analysis in the OCR we do the text recognition and page layout analysis we analysis the text blocks, text lines, paragraphs.
- Graphical processing in this we contain graphical data like halftone images, drawing, image, table and process this data. It further divided into the line processing and region symbol processing. Line processing for straight lines and region processing for the filled region.

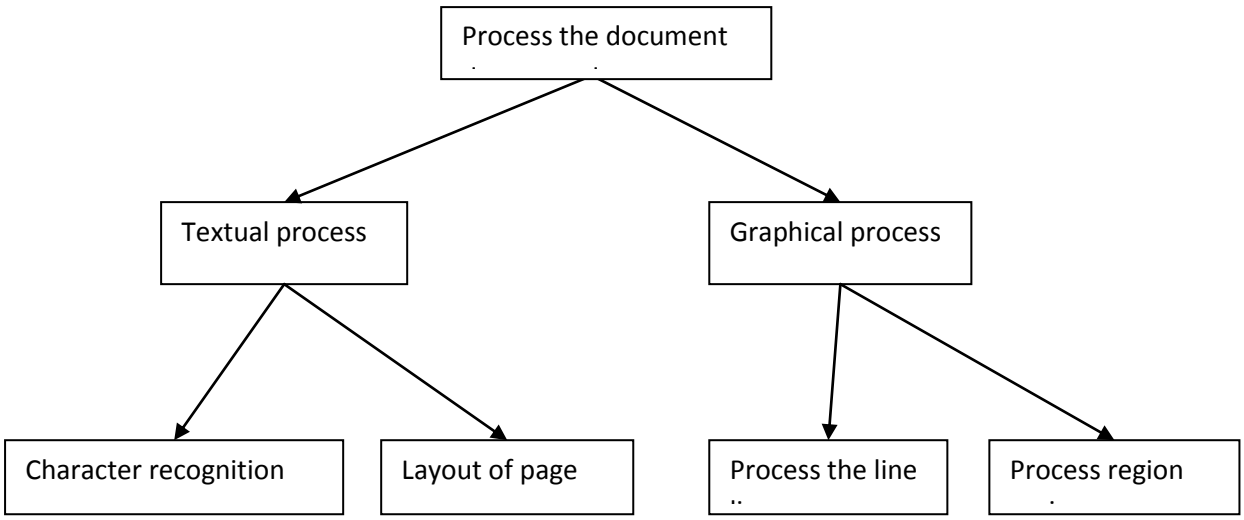


Figure 1.5: Document Image Analysis

1.4.1 Importance of Document Page Segmentation in DIA

One thing we notice in the DIA is that it process the document into two parts textual processing and graphical processing because we do not apply the same recognition on textual and graphical data so to recognition the document first of all we have to segment the document into textual and non textual parts which are basically done by the document page segmentation.

- Here we give u the brief introduction how the page segmentation work Image binarization it change the image into 0&1 format the pixel contain character give the 0 and non text part give the value 1.
- Noise removal in this part we remove the noise means remove the pixel which contain the more or less value as compare to the predefined value.
- Text and non textual separation by the classifier
- Detect the text region from the document by check the homogenous region.
- Text line detection is to separate the two text lines in the text region.

1.4.2 Overview of Steps Followed in Document Page Segmentation

Binarization of the image: In this process we change the gray scale image into the binary image. In our case we give the 0 pixel value to the textual part and 1 value to the non-textual and background part.

I. **Connected component analysis:** In this process we have done the connected component labeling. In the process of connected component labeling we give the same labels to that region which is belonging to each other. For example if the textual part is connected to each other then we give the same label to that region and then extract the connected component from it.

II. **Noise removal:** Noise removal is process to remove the noise from the image and that noise can be into the any form like speckles, dust etc. To remove the noise we can remove that pixel which has the value less then predefined value.

III. **Text and graphic separation:** In this process we are going to separate the textual and graphical component. After this step graphical component like table,

pictures, images are extract from the textual part. This step is the important step of the document segmentation. To see the importance of this step there are various type of method are develop to perform this step accurately. Here are some brief introductions of these methods.

- **Connected component analysis method:** In this method we are going to analysis that which component is connected to each other and after the analysis of this component we are going to extract these components. This method are very old method but still in use. This method has also some drawbacks like this method is unable to extract broken character because in broken character components are not connected to each other. The other disadvantage of this method is that it is unable to extract the small strings of character from the document.
- **Region based method:** Connected component approach are failed to extract the text component from the paragraph when the text component has the different color. So to overcome this problem we use the method which is applicable to extract the region of the text not only the connected component. Depend upon the shape of the text paragraph we can change the shape of the region.

There is also some drawback of region based method. This method is failed to extract the lines from the table which table contain the text component as well.

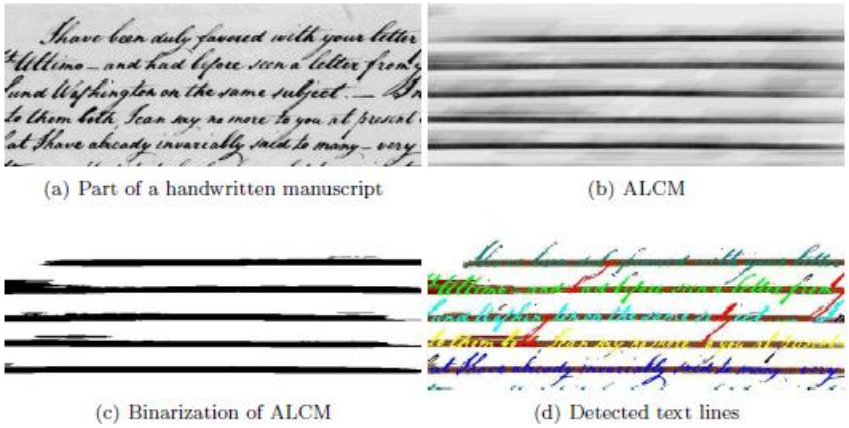
I. **Text region detection:** This method is used to detect the homogeneous region of text from the document it also separate the side note from the text. There is also the different method for extraction of the text region.

- Run length smearing algorithm: - This is one of the oldest algorithm is to detect the text region from the document. In this method we scan the document in both the direction horizontally and vertically on the predefined threshold value and make the homogeneous text block but the problem with this algorithm is that it is very difficult to select a particular threshold value.
- **Whitespace analysis:** This is one of the techniques to do the document layout analysis. With the help of this technique we are going to find out where are text blocks and columns are present in the document image.

II. **Text line detection:** In this process text line are extract from the document. In this we segment the text line from the text paragraph and segment that character properly which are touch to the other text line. There are various problems in text line detection like highly cursive text line, highly skew text line, gap between the text lines. So to overcome this problem we have also some techniques of text line detection.

- **Information energy technique:** This technique is mostly used in the handwritten text line detection. In this technique we are going to check the energy of each text line. If there is some energy present between text lines then we are not going to segment that line and if there is not any energy between text lines means there is not any character present between the text lines and we are going to segment that text line.

  As we show below we take the handwritten data and check there energy information value and change that value to the binary form and detect the text line from it.



(a) Part of a handwritten manuscript (b) ALCM

(c) Binarization of ALCM (d) Detected text lines

Figure 1.6: Check information energy of text

Chapter 2

Review of Literature

P.barlas et al., (2014) present the paper on document image analysis to extract the homogenous typed and handwritten text and successfully done the text/ non text segmentation and typed and handwritten text segmentation followed by the block segmentation to detect the white rectangle. This approach is applied on document of MOURDOUR COMPAIGN.

In the previous method [1] we apply the segmentation one type of handwritten language with less variability in the document structure but when we apply this technique on MOURDOUR COMPAIGN dataset which contain the handwritten text of Multilanguage like Arabic, English, French then this technique not give the good result.

In the present work we extract the homogenous block of handwritten and printed text with the help of connected component and block segmentation with white zone. In the first step we do the noise removal and remove the small and large connected component which are close to the border of the document and then the next step is to classification of text and non text with the help of predefined shape of connected component and then it further give to the multiple layer preceptor classifier to classify this and the next step is the layer separation in which we done the classification of printed text and handwritten text with the help of codebook in which we first of all construct the fragment and classify it by the multiple layer preceptor classifier .the next step is block segmentation in which we generate the block which contain the homogenous region with help of run length smearing algorithm and after this segment the document with white space.

This method achieves very good accuracy to generate the text block with the help of block segmentation and white rectangle analysis.

Limitation of this method is the miss classification of small connected component in the text layer.

Coston-anton boiangiu et al., (2014) proposed the algorithm to identify the text line by using information energy technique because for correctly recognition we have correctly segment the text line.

The previous work input image taken as the format of gray scale which is further taken as a binary or white and black pixel method by the previous defined threshold method. The pixel grey value highest from threshold value taken as black pixel and other one taken as white pixel but this method works well on printed document but in handwritten document we use the information energy technique .

In present algorithm energy map show the amount of information associated with each pixel if the pixel contain the high information value it means it contain the text if the pixel information value is low than it mean it show the space between text line. First step of this algorithm to calculate the energy map the pixel value and for accurately segmentation the direction of each text line also taken as consideration.

This algorithm shows good accuracy in printed as well as hand written document .the other advantage of this algorithm it check the value the energy information value of each pixel and

Also show the good result in high skew data.

Limitation of this algorithm it show less accuracy in more complex document format like where lower line character touch with upper line character.

Burcu yiddiz et al., (2013) developed many heuristic which together recognize and decompose the table which are present in the PDF and extract the data form it in the format of xml for easier reuse.

We define the existing approach into three type 1.predefined layout approach in this approach we use the template and scan the portion of the document if it fit into the template it is recognize as table but the limitation of this approach there has not two many template present.2.heuristic based approach in this approach we use the set of rule to define the table.3.statistical approach this is training based approach.

In this approach the first sub task is table extraction but this is not the easy task because each table has different format. We apply the PDFTOHMAL we had to extract the table information from an xml document with text element describe the position of text chunk

in a PDF file. There are two type of heuristics on the table one is table recognition in which we deal with xml document which not marks up table and we have to identify a portion of text as a table and the second one is table decomposition to correct identification of the header element so that that the correct assigning of data cells to header element.

This approach is very best for the table recognition task and properly finds the text position in the table.

The main limitation of this paper is it totally depends upon the output of PDFTOHMAL tool if it gives the wrong information than the user can do nothing with this. There is two type of table is use for experiment lucid and complex table and this approach is best only for the lucid table.

For the future work we use more statistical technique for utilize regularities for better result.

Ankush Gautam et al., (2013) employ the system wavelet and 2 mean classification for extract the text from image and in post processing step using the morphological operation like erosion and dilation.

There are various approaches regarding the text segmentation from the image like region based and texture based approach but this approach is failed when we apply the segmentation in complex document.

We start our approach to convert the RGB image into grey scale image

Grey (i, j) = 0.59 green (i, j) + 0.30 red (i, j) + 0.11 blue (i, j)

After this we apply the wavelet transformation which image decomposed into multi resolution frame in which every portion has distinct frequency and spatial properties. The image will be large so we apply the block processing which decompose the picture according to user specification and also resembles each block result into output image. After this we apply the k-mean clustering approach where k is for the input parameter in this approach we use the 2 mean clustering so we take the input parameter as a black pixel and white pixel. In post processing step we used morphological operation one is dilation to add the pixel to the object boundaries and erosion for remove the pixel from

object boundaries the number of pixel add or remove from the image is depend upon the structuring element to process the image .

This proposed method is very efficient for detecting the all kind of text and graphs from the real life document.

The limitation of this method is failed when there is very complex layout structure.

In the future work we improve the segmentation by segment the text from graphic image.

Priyadharshini N et al., (2013) propose a new technique of document segmentation into text image, drawing and tables and further document image is divided into blocks using run length smearing rule. DISCIPULUS tool used to construct genetic programming based classifier model which classified the document with 97.5% accuracy.

Till now there is lot of approach is use for document segmentation like top down approach in which we divide the document into smallest region which cannot further divided for ex run length smearing algorithm in bottom up approach pixel merge into character and character merge into word so on this approach is applied until whole document will be merged like run length smoothing algorithm and the third approach is the hybrid approach which is combination of both the approach like texture based method and GABOR filter.

In this approach the first step is preprocessing step in which we done the scanning binarization and noise removal and in the next step run length smearing algorithm divide the document into particular block where each block contain homogenous data the run length algorithm change the input binarization image x into output y image by

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.

Next step to give the unique label to each component this is done for to differentiate the each block. Further the feature extraction process gives the feature of the each block like their height, width so that it is easy to classify the text-non text region. After this process we applied the genetic programming for the machine learning classification to classify the text as text and non text as non text.

In this paper we use the genetic programming language which shows the highest accuracy as compare to the other classification algorithm. DISCIPULUS is world fastest genetic programming and analysis tool and for segment the document which contains multiclass

data one by one classification technique is used which classify the data in six multi classifier like text/image, text/drawing, text/table, table /image, table/drawing.

D.Sassirekha et al., (2012) present the two techniques under block classification methods; two methods are enhanced and evaluated for extract the text object from PDF image. The two techniques which were used in block based segmentation is Ac coefficient based technique and histogram based technique. In this approach we first change the PDF into image and then apply image segmentation on this.

In the previous work mostly we mostly use the block segmentation approach in this approach we divide the document into their particular region which contain the homogenous data but in the proposed method we discuss the two sub section of the block segmentation AC coefficient technique and histogram based technique.

Ac coefficient introduce discrete cosine transformation to segment the image into three parts background block which is the smooth regions and text and graphics block are high edge region and the image is the non smooth part of the PDF image. For the background region AC energy is calculated if it is lesser then the defined threshold value then it is calculated as smooth region else non smooth region. For further identify the image and text region there are two feature vector d1, d2 calculated. The work of d1 is to calculate the encoding length of text and non text region because the code lengths of text block is high due to high level of frequency content in this. The work of d2 is to measure how close a block to two colored block for two color projection we use the k-means clustering algorithm so the high contrast block has high chance to classify as text block. The second technique is histogram based technique in there segment there are series of decision rule from highest priority block type to lowest priority. After apply this two technology we estimate the text region.

This both technique show the accuracy around 92% .if user is willing little time for better accuracy then ac coefficient technique is better but if user tolerate slightly less reliable data then histogram technique is better.

The limitation of histogram technique is that it give the less reliable data and ac coefficient give the low retrieval time so that this technology both depend on each other.

In the future work we make more accurate result by relate these two techniques with each other.

Fattahc Zirari et al., (2012) this paper proposed a method of text and non text classification in document using graph based modeling and structural analysis.

There is three kind approaches is applied for document image segmentation 1.Top down approach in this technique we segment the document into smallest region until it not subdivide further the most important method in this approach like projection method but the limitation of this approach is that it require the prior knowledge about the document layout.2.bottom up approach we start from the smallest part of image called pixel and merge the pixel. The most important method which use this technique like RLSA, region growing method.3.Hybrid approach it is the mixture of top down and bottom up approach and merge the advantage of both this approach.

The first step in this proposed algorithm to find the homogenous region of all document with the help of minimum spanning tree and then that region categorized into textual and non textual part by structured analysis. In this approach we take the image as a undirected graph in which each pixel is changed by node and edge represent the relationship between the nodes and balanced by the sum of intensities of the pixel at ends. To measure the homogeneity of intensities we use the concept of internal difference. a graph is constructed over the entire image with each pixel pix being its own region and region are merged by analysis the edge in sorted order by increasing weight and check whether the edge weight is smaller than the internal Variation of both regions incident to the edge if true the region merged an internal variation of compound region is updated. After contain the homogenous region the second step to divide into text and non text part in their first step we calculate the textual part and filter out the majority of non textual part by calculate the histogram of the frequencies of the component size the components belonging to most significant peaks of this histogram are retained. The second step is eliminating the frequent noise component and the graphical one which are not in done first step in this step we do the segmentation on textual alignment which are mostly common for textual part.

This algorithm shows the great accuracy in segmentation of textual part. it show about the 98.5%of accuracy for text detection. And we also obtain the 97%accuracy to the detection

of non textual part the 3% error rate because some non text part has alignment same as textual part

Syed Saqib Bukhari et al., (2011) proposed an algorithm which done the modification in the Bloomberg algorithm which were used text/image segmentation.

Limitations of the previous Bloomberg algorithm is that it is use only to segment the text and halftone image but unable to segment the graphical image like drawing ,graph ,map etc. in the Bloomberg algorithm work on threshold reduction and basic morphological operation .in this we take the pixel value as 1 and 0 and apply the 4*1 threshold reduction on this means we some up the four pixel and if the value of this sum is greater than the threshold value then it is taken as 1 else 0 after apply this process we obtain the seed image and then to make the full halftone image from seed image them we do the 1*4 expansion so it can be equal to the input image.

In this approach we done the two modification in first modification we fill the hole the drawing images before the threshold reduction after apply the first modification it can done the segmentation of halftone as well as drawing ,graph like non text mages but there is some limitation like broken drawing line so to overcome this problem we done the second modification by reconstruction of broken drawing lines this is done by using Gaussian filter approach both on horizontal and vertical lines.

First of all Bloomberg algorithm is only capable to segment the text and halftone images but after the two modifications it show the great accuracy halftone image segmentation as well as drawing, graphs, map. To check the accuracy of this modify algorithm we test it on two type of document UNLV and circuits which show the 88.45% and 93.49% of corresponding. The limitation of this algorithm is that it is not best to the detection of non textual part.

In the future work we apply this algorithm on more complex document and also improve the performance of non textual detection.

Jayant Kumar et al., (2011) propose the method to extract the printed and handwritten text zone from the noisy document. In this approach construct the shape codebook of both handwritten and printed text and with the help of support vector machine we classify each this zone.

There are various approach applied for the printed text and handwritten text on the various level text block level, text line based, word level, character level .the other method on classification like text block level by check the feature and layout of paragraph of text and the other one is classification by text line but this algorithm show the accuracy only for the English language but in this method we talk about the classification of printed and handwritten Arabic language and also this method is applied where printed and handwritten text coming side by side.

In this approach first of all we segment the document into zone and then we manually select the printed and handwritten text zone and construct the shape codebook and the behalf of shape codebook we can train the Support vector machine classifier which further classify the printed and handwritten text.

This approach is very good approach for classification printed and handwritten text from the document which contain the many type of content. We apply this approach without the segmentation of word and text line segmentation.

In the future work we can do the further work in construct this method for the other kind of language as well.

Loudloudis .G,et.al(2009) this paper proposed the method to segment the document into text lines and words. Text line segmentation achieved by apply the Hough transform on connected components and create the false alarm which tell that which text line are not correctly segment by the Hough transform and on that line we apply the novel method based on technique of skeletonization. Word segmentation is done in the two steps one is compute the distance of adjacent component and second step done the classification of previously calculated distance.

For the text line segmentation this algorithm is done the comparison with the fuzzy RLSA and projection profile and for word segmentation we comparison this algorithm with RLSA and projection and we test this algorithm on ICDAR datasets and this algorithm show the positive result.

This method improves the accuracy of previously defined text line segmentation and word segmentation methods. In the future use the punctuation detection method and feedback from character segmentation and recognition module.

Shafait Faisal,et.al (2006) This paper show the comparison between the six algorithm of document page segmentation and check there performance for both the default and optimized parameter on UW-III collection and we find that out of all six algorithm there is not any single algorithm perform the best out of other.

Out of six algorithms one algorithm is the dummy algorithm. This algorithm take the whole document as a single segment and check that what we can achieve without do anything. The second algorithm X-Y cut takes the whole document as root node and segment the document at the leaf node. This process is start until whole document cannot be segmented. The third algorithm is the run length smearing algorithm in this we take the image as a binary image where 0 represent the white pixel and 1 represent the black pixel and this algorithm work on the two concepts on is change the 0 pixel of image x to 1 pixel of image y if the value of the 0 pixel is less than the predefined value and 1 in x is unchanged in y. The fourth algorithm is the white space analysis. In the first step of this algorithm we check the white space present in the document and in the second step we are going to combine these white spaces. Constraint text line detection we count the number of white rectangle present in the document and after combined this spaces we are going to find the text line in it. The Docstrum algorithm is based on bottom up approach in this approach we combine the connected component and make the cluster from this. The sixth algorithm voronoi diagram also based on bottom up approach. In this we join the sample point from connected component by using the simple rate.

Each algorithm shows their accuracy in different field and show the good accuracy. The reason of good accuracy is also because of same resolution of UW-III paper.

In the future work we can check the accuracy of this algorithm on more heterogeneous datasets.

Strouthopoulos.c,**et.al(2001)** This paper proposed the method to extract the text from the complex color document which contain the drawings, graphs and thousands of colors. This paper does the combination of two techniques ACR and PLA. The ACR technique is to find out optimal number of color and change the document into principal of them. According to the principal color we obtain the color plains and after this PLA technique is applied to the each color plain to obtain the text region and after this merging process is applied to combine the text regions which are obtain from the color plains and produce the final document. This proposed method is applied on to the various kind of color document and each document has the resolution around 200dpi.

Chapter 3

Present work

3.1 PROBLEM FORMULATION

Document image analysis the process to analyze the structure as well as textual and graphical component of the documents. The main purpose behind this is that so that computer can understand the document as human does. There are lots document stored in computer daily so it is important that computer can understand the document as it understand the other media. Document image analysis is the technique which can help the computer to understand the layout of the document and extract the useful from it. In our problem formulation we take the dataset of handwritten bill which are not basically use by the computer friendly people. So it is important to analyze this kind of dataset and extract the useful knowledge from it. The main problem of this kind of dataset is that it contain the heterogeneous type of data means it contain the printed text, handwritten text and graphical data. In this research work we are going to segment all this heterogeneous data and extract the handwritten data from it and after we are going to recognize this handwritten and map it in structured format so to make this problem user specific.

3.2 Objective

Document segmentation is the preprocessing step of document image analysis .It play very important role because without we never can recognize the document image. But there are lots of works to be done in document segmentation and there is some objective to work on document segmentation.

- Segment the documents which contain the heterogeneous component.
- To analyze the various document segmentation technique.
- To propose the enhancement in document segmentation technique to improve the accuracy.
- Extract the handwritten text from the handwritten bill and map it onto the text document.
- Provide the user specific approach.

3.3 RESEARCH METHODOLOGY

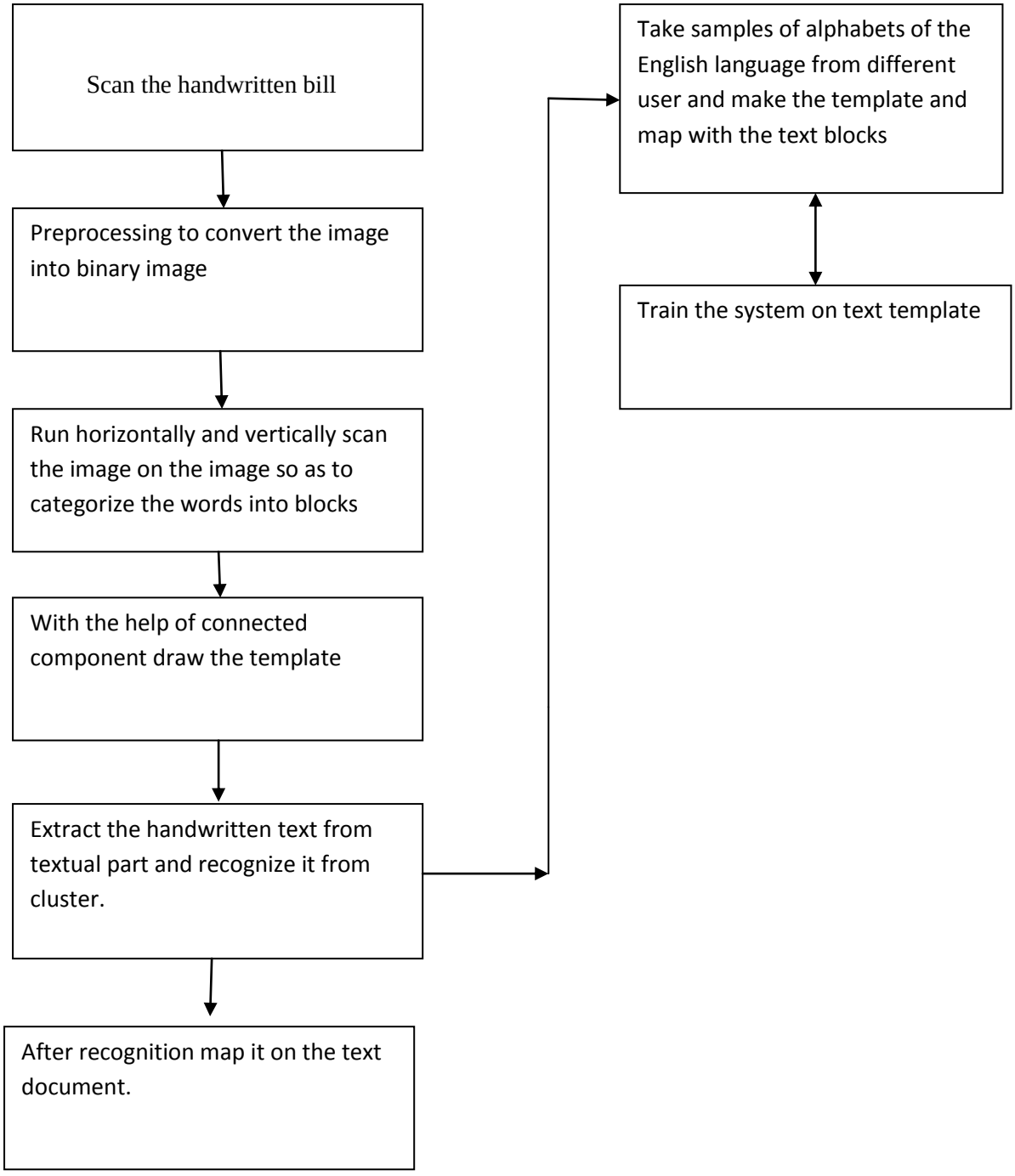


Figure 3.1: Flowchart

In our research methodology we are following these steps:

Step 1: Train the system

First of all take the samples of alphabets of English language from the many users and make the template of each individual character to map with segmented block text. We map the individual word of the text block with trained system and recognize the each text block.

Step2: Scan the handwritten bill

The second step is to scan handwritten bill which is in the form of bill which contains the heterogeneous type of content like printed text, handwritten text and graphical component like table. In our research we take the handwritten bill which contains the all kind of heterogeneous kind of data. It contains the table which is in the graphic form and it contains the both kind of text printed text and handwritten text and it include the text image.

Step3: Preprocessing

The third step is preprocessing step in which the scanned image converted into we first of change the image into grey scale image we change the image into grey by taking the threshold value if the pixel exceed that pixel value then it is count as white otherwise else black. After this we convert that grey scale image into the binary image in the binarization process we change the image into 0&1 form.

To change an image to binary image in MATLAB we simply use the two or three commands. For example take any image name ‘bill’ to change this image to binary image we use the following command like ‘i=im2bw(‘bill.png’)

Step4: Segmentation of the words

In this step used the technique of connected component which are going to segment that component which are connected to each other and make the bounding box around the connected component according to their properties and segment each word or character. This approach also used the feature based approach because in this method we only extract the handwritten text from the heterogeneous data which include the printed text as well as graphical component.

Step 5: Draw the template

In the fifth step we can draw the template of the table and extract the text region part from the table with the help of connected component approach. In this approach each pixel check the value of their neighborhood pixel if the neighborhood pixel value is the same or nearby value then it taken it into the same region. By following this process we can easily draw the template which contains the table because in the line the each pixel value is same but in the text value is changing pixel by pixel. So on the basis of this feature extraction we detect the text and non text part and give it to the MLP classifier to classify this document.

Step 6: Recognition of handwritten text block

The next step is to recognize to the each block of the text with the predefined template. In this step each handwritten text block are recognized and further this text are mapped to the text document

Our last step is post processing step in which we done the segmentation and the recognition of handwritten text line and map it on to the text document.

Step 7: Map to text document

After recognition recognize text map on to the text document.

Chapter4

Results and Discussions

In the proposed method we are going to analyze the handwritten bill dataset which contain the heterogeneous kind of data like handwritten text, printed text and graphical component. In this method we are going to segment this handwritten bill and extract the handwritten data from it and map it into the structured form.

Figure 4.1: Handwritten English character samples

As the above figure 4.1 show we are going to train the neural system with handwritten data of different user. In this following diagram we are take the sample of possible combination of a character.

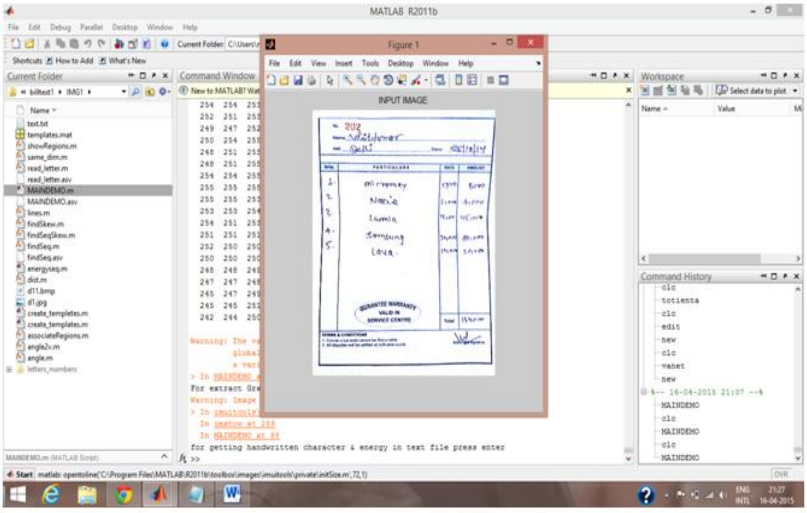


Figure 4.2: Scan the document

As the above figure 4.2 show that we are going to scan this handwritten bill for further

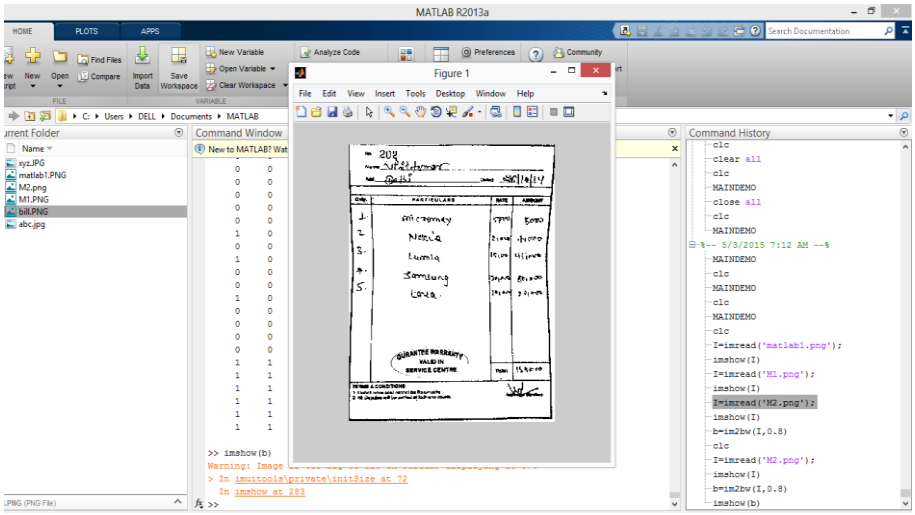


Figure 4.3: Change image into binary form

In the next step we are going to change the image into binary form which contains the two kind of pixel 0&1. Figure 4.3 show the template in binary form

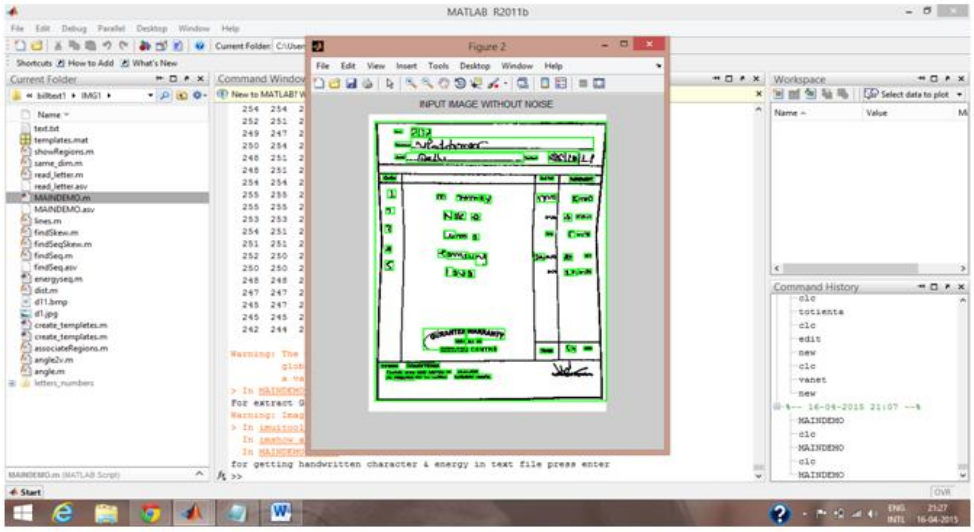


Figure 4.4: Segment the textual words

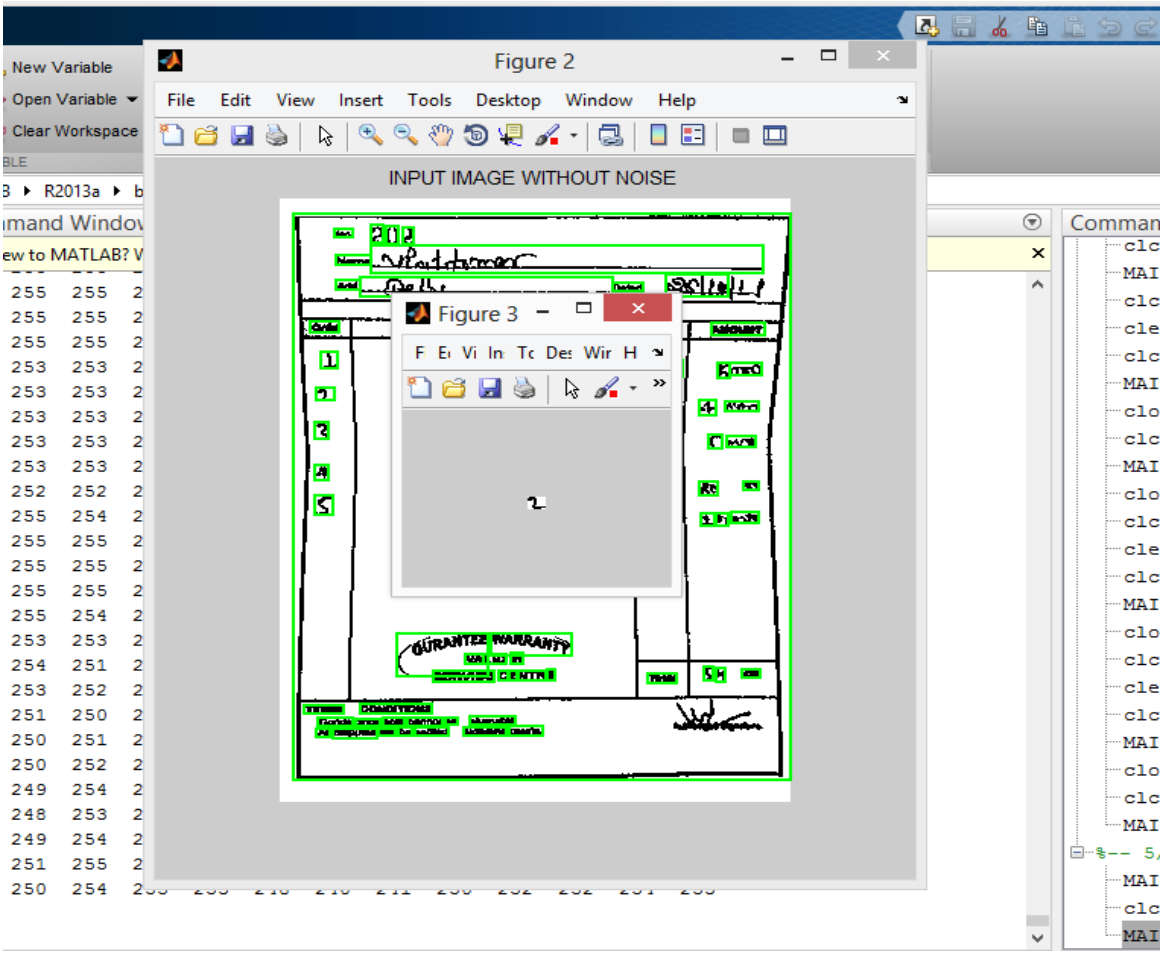


Figure 4.5: Processing of the word

In the above figure 4.4 and 4.5 we are going to show that how we are going to segment the each and every word and how we are going to process each and every word.

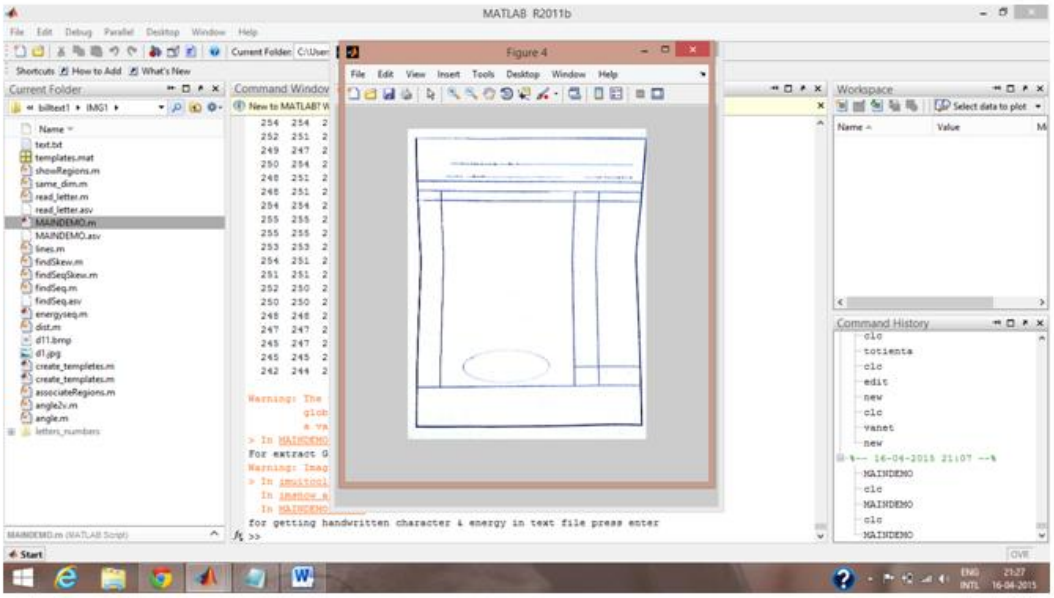


Figure 4.6: Extract the bill template

As the above figure shows that we are going to extract the bill template from the handwritten bill. To extract the graphical component we train the classifier which shows the accuracy around 80 -82%.

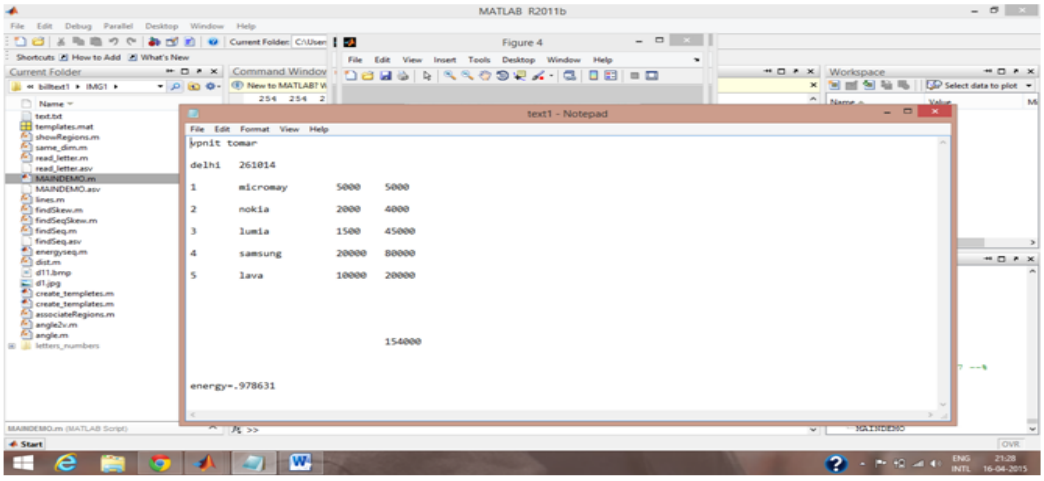


Figure 4.7: Recognition of handwritten text

After the extraction of the graphical component we are recognize only the handwritten text from the bill template. For the recognition of the handwritten text we are going to map this handwritten text with train neural network. Accuracy of our recognition system is around to 93.4%.

Chapter 5
Conclusion & Future Work

In this purpose method we enhance the technique of document segmentation which contains the heterogeneous component and we apply this segmentation technique on the dataset of shopkeeper bills. In this method we segment the bills which contain the heterogeneous type of component like handwritten text, printed text and the graphical component and segment this component and extract the handwritten text and recognize it with the help of neural network and map it to the text document.

In the future work we can map this text on to the excel sheet or in the more structured form and use another data set to map on excel sheet which are appropriate to the excel sheet and also more accurately improve the graphical extraction method.

Chapter 6
References

I. Books

James Allen (2005) natural language understanding, pearson education Singapore.

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal(2010) natural language processing, eastern economy edition New Delhi.

Lawrence O'Gorman,Rangachar Kasturi(1997)document image analysis

II. Research papers

P.Barlas, S.Adam, C Chatelaine and T Paquet (2014) "A typed and handwritten text block segmentation system for heterogeneous and complex document", publish in document analysis system, france(2014).

C.A Boiangiu, R.Laonitescu,M.CTanase [2014] "handwritten document text line segmentation based on information energy"publish in int j comput comm.,issn 1841-9836.

Brucu Yidiz, katharina Kaiser and Silvia Miksch[2013] "A method to extract the table from PDF files"

Ankush Gautam[2013] "segmentation of text from image document",publish in international journal of computer science and information, Vol,4(3), 2013,538-540.

Priyadharshini N, MS VIjaya "genetic programming for document segmentation and region classification using Discipulas", international journal of Advanced research intelligence,Vol.2,No.2,2013.

D Sasirrekha ,Dr.Chandra "enhanced technique for pdf image segmentation and text extraction", international journal of computer science and information security,Vol.10,No.9,2012.

Fattahc Zirari, Driss Mammes ,AbdellatiifiEnnaji and stephane Nicolas[2012] " A graph based approach for heterogeneous document segmentation"springer velag berlin Heidelberg ,424-431,2012.

S.S Bukhari ,Faisal Shafait andBreuel Thomas M [2011] “Improved document image segmentation algorithm using multiresolution morphology” Document Recognition and Retrieval XVIII, SPIE, 7874:1–10, 2011.

JeevanKumar,Rohit Prasad ,Huiaigu Cao, Wael Abd-Almageed ,David Doermann, PremkumarNatarajan“shape codebook based handwritten and machine printed text zone extraction” Document Recognition and Retrieval,SPIE7874,2011.

G.Louloudis,B.Gatos,I.Pratikakis,C.Halatsis[2009] “text line and word segmentation of handwritten documents” Elsevier pattern recognition 42(2009)3169-3183

Faisal Shafait, Daniel Keysers, and Thomas M. Breuel[2006] “ Performance Comparison of Six Algorithms for Page Segmentation” LNCS 3872, pp. 368–379 Springer-Verlag Berlin Heidelberg 2006

C. Strouthopoulos, N. Papamarkos, A.E. Atsalakis[2001] “Text extraction in complex color documents” Elsevier Pattern Recognition 35 (2002) 1743–1758

.

Chapter 8

Appendix

DIA: - Document image analysis

OCR: - Optical character recognition

RLSA: - Run length smearing algorithm

ACR: - Adaptive color reduction

PLA: - Page layout analysis

RGB: - Red green black

MLP: - Multi Layer Perceptron

NLP: - Natural language processing

PDF: - Portable document format