IOSR Journal of Engineering
May. 2012, Vol. 2(5) pp: 1029-1031

Removal of Graphics from Text-document and Segmentation of Gujarati Documents Using Connected Component Theory

Himanshukumar U. Prajapati [1], Prof. S. Ramamohan[2], Ms Sonam Chauhan [3]

[1](Faculty Of Tech & Engg., MSU,Vadodara)
[2](Chief coordinator,MCIT-OCR Project,MSU,Vadodara)
[3](Electrical Engg. Dept.,Faculty Of Tech & Engg., MSU,Vadodara)

ABSTRACT
Very little work is found in the literature for recognition of Indian language scripts. We have work on graphics separation & text segmentation for Gujarati language documents.
Text/graphics separation aims at segmenting the document into two layers: a layer assumed to contain text and a layer containing graphical objects. In this paper, we propose method of choosing right threshold to extract text from graphics. We consider connected component characteristics to differentiate between text and graphics. It is based on the observation that connected components of graphics are comparatively larger than text. Experimental results show that while removing graphics, no damage on connected components of text in the image. Segmentation is one of the most difficult tasks in digital image processing. In which each character of the document has to be identified individually.

Keywords **-** connected components, graphics separation, minimum bounding box (MBB), and segmentation.

INTRODUCTION
OCR software for Gujarati Language is almost in final stage of development after the continuous hard work for many years but there are several challenges which have to be solved.

***Challenges*:-** Gujarati OCR has been currently in development but faces many challenges that reduce its efficiency.-In some kind of books, there are characters having irregular thickness. Hence, problem occurs due to Broken and merged characters in document images pose serious challenges for recognition accuracy.If these problems can be reduced by any means, then overall efficiency can be increase by up to 10-15% a now

-There are graphics related issues; Which reduces this system's efficiency.

-Recognition of handwritten document is difficult one, because different persons have their own writing style. New systems have been developed for that which is known as ICRS.

-There are problems of fusion of different languages includes in single document Like: English & Gujarati or Sanskrit & Gujarati

The document image generated as input to OCR may be subjected to various operations to remove the artifacts and enhance the image quality so that subsequent processes give expected results. Thresholding (or binarization) & graphics removal are complex challenges during pre-processing stage. Documents containing graphics is one of the challenges to develop OCR (Optical character recognition system). It reduces OCR efficiency rapidly. Graphics has not uniform shape or design, it always varies. So in our database we have no data about any particular graphics. That is why it is necessary to remove graphics for better result of OCR, output of which is editable text.

II. CHARACTRISTICS OF GUJARATI SCRIPT
As other Indian languages the character set of Gujarati comprises of 36 consonants, 12 vowel and 6 signs, 12 dependent vowel signs, 10 digits. The consonants can be combined with the vowels and can form compound characters as shown in Fig 1.

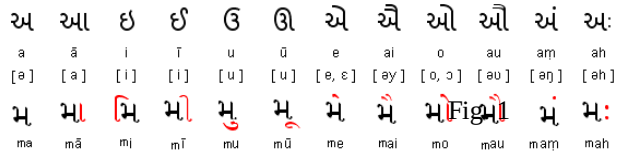


Fig.1

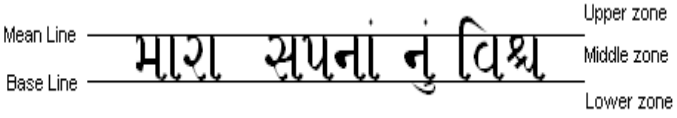
A word may be formed by combining the basic character(s), which may by combined with vowel(s). Sometimes the conjunct consonants may also form part of the word. Collection of words will form a line and collection of lines will form text.

Fig. 2 shows a line of Gujarati text, which can be considered as being logically divided into three horizontal parallel lines [3].The base line, the imaginary line separating the middle and lower zone on which consonants and independent vowels are written (Middle zone).

The line below the base line, used for writing dependent (lower) vowels (Lower Zone).

The line above the mean line, used for writing dependent (upper) vowels (Upper zone).

A typical example of zoning is shown in Fig



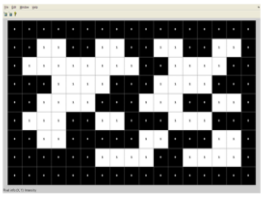
Gujarati Script from work Perspective

Consonants
ક ખ ગ ઘ ઙ ચ છ જ ઝ ઞ ટ ઠ ડ ઢ ણ ત થ દ ધ ન પ ફ બ
ભ મ ય ર લ વ શ ષ સ હ ળ ક્ષ જ્ઞ
Vowels
અ આ ઇ ઈ ઉ ઊ એ ઐ ઐ ઓ ઔ ઔ અં
Some Conjuncts
ક્ક ક્ય ક્લ ચ્છ દ્ર ત્ર સ્પ સ્ત ર્વ ક્ક ત્ય ત્યં ત્ત દ્ર
Consonant – vowel
જા ગી જી બુ ખૂ ઠે કુ કે પૈ પો ડો કૌ કં
Conjunct -vowel
જ્જા કે ક્લે ચ્છિ દ્રી ત્રુ સ્ન્
Vowels modifiers
ા િ ી ુ ૂ ૃ ે ૈ ો ૌ ્ ં ઃ ઁ
Numerals
૧ ૨ ૩ ૪ ૫ ૬ ૭ ૮ ૯ ૦

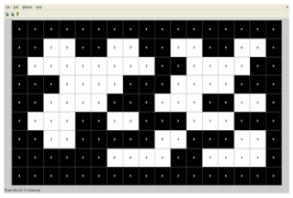
While trying to remove graphics from an image, we had to consider some parameters like consonants, vowels, conjuncts, symbols and their combitions. Other from Gujarati Script it should identify as graphics.

III. CONNECTED COMPONENT LABELING

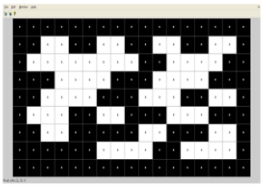
Connected component labeling (alternatively connected component analysis) is an algorithmic application of graph theory, where subsets of connected components are uniquely labeled based on a given heuristic. Some of the graph theoretic concepts used in this in this algorithm are Adjacency & Connectivity. These concepts are defined in the context of digital images [4].



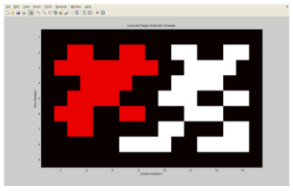
[The array from which connected regions are to be extracted]



[After the first pass]



[After second pass]



[two different regions that have been found]

connected component characteristics to differentiate between text and graphics. We decide threshold to remove large graphics from the image. This is the major cause of reduction of OCR efficiency. First we convert an image into a binary image. Binarized image is the output of thersholding method. We use Otsu[4] method to decide threshold value for binarization.

Graphics Removal

Algorithm for binarized image as input is as below:

- Labeling the binarized image
- Find the max number of labels
- Find the number of connected component for each label
- Make them in ascending order
- Set the threshold value at that label above which there is rapid change after consecutive difference
- Remove label which has number of connected components above threshold limit, which is graphics

B. Segmentation

We have to identify each character individually. Once this is done we find top four co-ordinates around the character. After that we make the MBR around the rectangle

Segmentation

Algorithm for segmentation method as follows:

- Convert image into binary image
- Labeling of the image
- Find the maximum number of labels in the image
- Sequence the labels in the image
- Find the co-ordinates for each label
- Make MBB around the label

Results for Segmentation:

Type Of Images	Number Of Characters Checked			Efficiency Of Segmentation
	Total	True	False	
Grade – A*	600	600	0	100 %
Grade – B**	600	592	8	98.66%
Grade – C***	600	588	12	98 %
Grade –D****	600	452	148	75.3%

In the results of segmentation, efficiency reduces only due to problem of broken and merge characters.

* means images having no broken or merge characters

** means images having least broken or merge characters

***means images having blur, having more broken or merge characters then B-type

**** means images having blur, low illumination, large number of broken or merge characters then C-type

V. CONCLUSION

In the current work, we have developed a fairly accurate text/graphics separation methodology then the method used for Gujarati Language, a now. Large graphics can be easily removed. This algorithm has limitation to remove small graphics as in result because while we trying to set lower threshold limit, the modifiers are also removed. This method also removes texts that connect to graphic. So, shows the limitation of graphics removal method. On other side segmentation methods is efficient and generate exact MBB around the character for extract. Problems are seen only on broken or merge characters where connected components not exactly represent any character or modifier.

VI. FUTURE SCOPE

Further moving on from graphics removal method, other method can be included in addition so that to remove small graphics. Segmentation method to deal with broken and merge character can be developed. That will overall increase the OCR's efficiency.

ACKNOWLEDGEMENT

This work has been carried out at MCIT-OCR centre at Linguistic Dept. Arts Faculty, MSU, Vadodara. I am so thankful to Ms Soram Kotia & Prof. A. I. Trivedi at Faculty of Tech. &. Engg. for their continuous guidance and precious help during work

REFERENCES

Journal Papers:

1) Ruini Cao, Chew Lim Tan School of Computing, "Text/Graphics Separation in Maps" National University of Singapore.3 Science Drive 2, Singapore 1175 (Text graphics separation) ,2001
2) Karl Tombre, Salvatore Tabbone, Loïc Pélissier, Bart Lamiroy, and Philippe Dosch, LORIA "Text/Graphics Separation Revisited" B.P. 239, 54506 Vanduvrelès-Nancy France (Text graphics separation), 2002
3) S.Rama Mohan, Jignesh Dholakia, Atul Negi. "Zone identification in the printed Gujarati Text" Processing of the 2005 Eight International Conference on Document Analysis & Recognition(ICDAR'05)

Books:

4) Gonzalez, R.C. and Woods, R.E,"Digital Image Processing", Second Edition, Pearson Education, Singapore

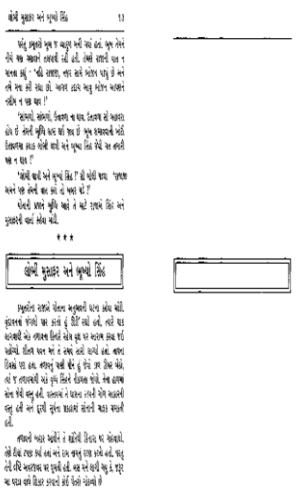
Theses:

5) Mr. Jignesh Dholakia, Mathematical Techniques for Gujarati Document Analysis and Character Recognitoin,2010

Websites:

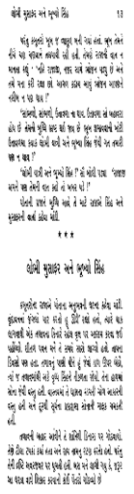
6) http://en.wikipedia.org/wiki/ Optical_character_recognition (Basics of OCR)
7) nstats.un.org/unsd/demographic/meetings/ wshops/new.../docs/05a.ppt(Basics of OCR)

Text-Graphics Separation Result



Input

Removed Graphics



Output

Text-Graphics Separation Result

Input

Removed Graphics

Output