Volume : 1 | Issue : 3 | December 2011 ISSN - 2249-555X

Research Paper Engineering

Analysis of Images of Printed Gujarati Documents for Semi-Automatic Ground Truth Generation

*Parmar Ranjitkumar K,

*Lecturer Electrical Engg Govt Polytechnic, Junagadh

ABSTRACT

Benchmarking of any product is necessary to evaluate the performance in all possible Conditions. As the name suggests, ground truth is a collection of facts in terms of the input data and expected output from the system when the input is subjected to the system to evaluate its Performance. Document Image Analysis is an integral part of most of the Optical Character Recognition Systems these days as they are supposed to take images of documents having almost any format as input and rearrange the recognized text in the same format as in the printed document.

In order to evaluate the performance of such a system , it is necessary to have a Collection of large number of images with the expected output when these images are given to our DIA system. Creating such a collection manually is a tiresome and time consuming process. Hence, it is necessary to have a system which can simplify the task by providing some reasonable output and then provide the tool to correct the output to match the expectations

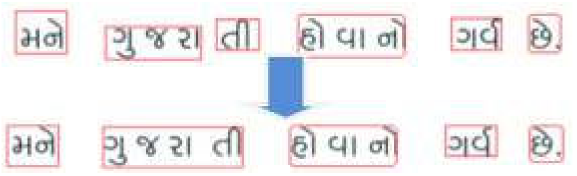

Fig. 1. Image illustration for semi automatic ground truth generation

Fig. 2. Screenshot GUI module showing original image, word box image along with word XML data

1. Introduction

Optical character recognition of Indian scripts is a mature technology now. Work is going on in various parts of world on different Indic scripts. In [1], Atul Negi et al describe various techniques for the South Indian Telugu script but very few work progressed in Gujarati. So far, only one documented effort for developing Gujarati OCR,viz. by Samir Antani [4] has been published. Here too some limitations in the approach were observed

1.1 Document Image Analysis (DIA) :

Document image analysis [3] is the study of converting documents from paper form to an electronic form that captures the information content of the document. Necessary processing Includes recognition of document layout (to determine reading order, and to distinguish text from diagrams), recognition of text (called Optical Character Recognition, OCR), and processing of diagrams and photographs. The processing of diagrams has been an active research area for several decades. The objective of document image analysis is to recognize the text and graphics components in images of documents, and to extract the intended information as a human would.

2. Gujarati Script and its problems:

Gujarati. Gujarati has 11 vowels and 34+2 consonants as shown in [1]. Apart from these basic characters set it gas 11 modifiers which are combined in various combination with the consonants or vowels. Unlike English Gujarati is having many other problems to be handled, because it is more complex script to write with. The major problems of this script that require special care are:

1. Attachment of modifiers before, after, above, below and within the basic consonant cluster.
2. Large Number of symbols.
3. Similarity of symbols.
4. Touching and broken characters.
5. Changes in Script.
6. Unavailability of language model

3. Necessity of Ground Truth Generation:

Ground truth is a representation of the agreed correct result of the ideal layout analysis method. Typically, training and evaluation require the ground truth data to be keyed in manually from the scanned image, but this is often a prohibitively labor intensive and error prone process. Furthermore, it may require domain experts, especially for processing multilingual documents. The present paper proposes a method for annotation of printed Gujarati text line and words. It is expected that this approach shall make the way smoother for the development of Gujarati OCR systems for the generation of Ground truth. In this paper, we have proposed a system for the quick generation of ground truth by making it semi automatic.

4. Semi automatic Ground Truth Generation:

It is necessary to have a system which can simplify the task by providing some reasonable output and then provide the tool to correct the output to match the expectations This alleviates burden to a great extent. This system takes Gujarati printed document as an input image and after certain basic preprocessing like RGB to Gray conversion, Binarization, Noise Filtration etc, it detects the text lines and words bounding boxes. That document is then acted upon by the DIA system for Ground truth Generation The DIA system will provide XML document which will be having data like top left and bottom right coordinates of various lines and words within the said document. System will draw bounding boxes on each word on Word Box Image and on each line on Line Box Image. Now it's the time for the visual inspection of the document . If any miss detection of line boundaries in case of Line Box Image or word boundaries in case of the Word Box Image is found then through the INTERFACE TOOL that has to be corrected. Now those bounding boxes which were corrected will have the different bounding box coordinates, So those old coordinates are updated in the data base with the latest true coordinates. Now this updated database is ready to be utilized as more corrected ground truth.

4.1 Algorithmic steps for semi automatic ground truth generation:
1. Load the Image.
2. Convert the image to Gray.
3. Binarize the Image.
4. Filter out the Noise from the image using Median filter.
5. Perform Horizontal Projection analysis for Lines detection.
6. Apply Vertical Projectionfor Gap analysis.
7. Calculate Word Gap and thus total number of words in a line.
8. Extract Lines coordinates.
9. Extract Words coordinates
10. Find fittest bounding boxes.
11. Draw Red Boxes on detected lines.
12. Draw Boxes on detected words.
13. Visually see for wrongly detected lines.
14. Apply Manual correction using mouse to have true lines boundary.
15. Update database in XML file format.
16. Visually see for wrongly detected words.
17. Apply Manual correction using mouse to have true words boundary.
18. Update database in XML file format.

4.3 Program Path:

4.3.1 Image is loaded using GUI prepared with the help of Qt Designer.

4.3.2 If input image is not Gray then it has to be converted to gray using equation:

Grey=(int)(0.33*red+0.56*green+0.11*blue)

4.3.3 Gray image is binarized using threshold comparison mechanism. If the concerned pixel's
intensity is more than threshold then change it to white else black.

4.4.4 Noise Filtration is done by Median filter, which arranges all eight neighbors in ascending order as per the intensity level and picks up fifth member as the median. Median filters are quite popular because, for certain types of random noise, they provide excellent noise reduction capabilities, with considerably less blurring than linear smoothing filter of same size.

4.4.4 Horizontal projection analysis:
Horizontal projection shows no. of pixels in each row. Thus successive blank rows indicates gap between two text lines. Thus, with this simple criteria starting and ending details of each text line can be achieved.

4.4.5 Vertical projection analysis:
This process is repeated on each erected text line with predefined starting and ending rows. The successive gaps in this projection suggests gap between two characters. Thus, from above two projection profiles each character's and text line's bounding details can be gathered.

4.4.6 Word gap analysis:
Gaps between characters will be much lesser than that between the words. To find out words, depending upon various practical trials, it has been found that following formula gives almost satisfactory results.

Word Gap = (Max . Gap + Average Gap) / 2.

4.4.7 After finding words and lines, their coordinates are transferred in an XML file format.

4.4.7 Fit bounding boxes are drawn around each line and word in to new images. These boxes are then visually inspected for wrong boundary identification.

4.4.8 Rubber banding boxes:
If correction required correct with mouse events. These newly redrawn bounding boxes coordinates are updated in the XML database. Now one can rely on this data set to compare the output of any OCR system for the performance verification.SOftwares used:Linux :Platform for developing and testing of source code.

Gnu C++ :Basic programming language.

Qt Designer :A GUI Toolkit for enhanced user friendly features.

XML :Extensible Markup Language for data representation and sharing

5. Conclusion

The aim of this project has been to develop a system which performs analysis on scanned images of printed Gujarati documents. Most of the processing during document analysis can be carried out without much assistance from the user; however stages need to be guided by the user for correct results. Though the user is not intended to be the part of the routine process, at times however, the system may ask for assistance. In these situations, user's interaction becomes helpful in making the system more reliable. Thus, the project carries out the document image analysis and generates an XML file, that efficiently and clearly represents more correctGround truth of printed document image having Gujarati text
.

6. Future Extensions:
1. Interface for correction of recognized regions and annotating the non recognized part can be added.
2. Zoom in & out functions to view the images can be included for better visualization.

REFERENCES

[1] Dholakia Jignesh, Negi Atul,Ramamohan S."Zone Identification in the printed Gujarati Text" Proc. Of 8th International Conference on Document Analysis & recognition 2005. | [2] Gonzalez,R.C. And Woods, R.E., Digital Image Processing, Second Edition, Pearson Education, Singapore. | [3] Bunjke H. & Wang P.S.P. "Hand book of character reorganization & Document Image Analysis" World Scientific,Singapore,1997.| [4] Samir Antani, Lalitha Agnihotri,"Gujarati Character Recognition", in proc. 6th ICDAR, pp.418-421,1999.