Text Recognition using Image Processing and Translation

Prof. A.K. Gaikwad[1], Mayur Pabalkar[2], Meeraj Ansari[3], Saket Kolte[4]
[1]*Guide,* [234]*Student*
[1234]*Department of Computer Engineering, Sinhgad Academy Of Engineering, Pune, India*

Abstract- Recognition of scene text is a challenging problem compared to the recognition of printed documents. An approach is proposed to recognize text in complex background natural scene, word formation from recognized text and word translation into user defined language. The proposed approach is robust to different kinds of text appearances, including font size, font style, color, and background. Combining the respective strengths of different complementary techniques and overcoming their shortcomings, the proposed method uses efficient character detection and localization technique and multiclass classifier to recognize the text accurately. The proposed approach successfully recognizes text on natural scene images and does not depend on a particular alphabet, text background. It works with a wide variety in size of characters and can handle up to 20 degree skewness efficiently.

Keywords- OCR, translation.

I. INTRODUCTION

Text recognition is a challenging and intricate process due to the often bad quality of images, different backgrounds or different fonts, colors, sizes of texts and hence there is still enough room of improvement in text recognition techniques which plays a vital role in improving the overall performance of the text recognition system. Scene images are characterized by complex background, perspective distortion, low resolution and poor quality. Scene text can appear with any slant, tilt, in any light and upon any surface and hence hard to detect, localize and recognize. There are a number of pre-processing steps followed by the actual recognition in the implementation of text recognition system.

The process of character recognition of any script can be broadly broken down into three stages; text extraction, classification, post-processing. Typical text extraction includes a collection of operations that apply successive transformations on an image. It takes in a raw image, removes noise, distortion, skewness and detect text segment by applying various segmentation & connected component analysis and feature extraction techniques. The selection of a stable and representative set of features is the heart of pattern recognition system design. Perhaps the most consequential one is the selection of the type and set of features among the different design issues involved in building a text recognition system. The classification stage is the main decision making stage of a text recognition system and uses the features extracted in the previous stage to identify the text segment according to preset rules.

The post-processing stage, which is the final stage, improves recognition by refining the decisions taken by the previous stage and recognizes words by using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies, lexicons, and other context information.

II. OVERVIEW

Many approaches have been proposed for text recognition in natural scenes which can be classified broadly to three categories. Firstly, approaches that can recognize the segmented text by proposing their own features with classifiers training which works well for specific languages and specific data , secondly approaches that can recognize and binaries the text without segmentation of text lines using multiple hypothesis frames work thirdly approaches that can improve recognition rate by enhancing the text through binarization which works well for carved text which may have complex background and low resolution ,for learning features from unlabeled data many approaches have been proposed . Authors in have shown that performance can grow with large numbers of low level features while in , authors have shown that performance can grow with large number of high level features.

III. MOTIVATION

While visiting foreign countries for tourism or for business meetings, people find it difficult to read & understand the local languages.

Using this methodology, it will become easier for them to read & understand the text written in native languages.

Objective

- To recognize text from natural images and scenes.
- To translate text into user defined language.

Literature Survey

1) Text Extraction from Document Images using Edge Information by Sachin Grover, Kushal Arora, Suman K. Mitra. Says In this paper, they have demonstrate that simple texture measures based on edge information provide very useful information for text detection from complex document images.

Disadvantages:

When the gradient of intensities of text and image are quite similar. Finding a generalized value which can work on every kind of image also needs some working.

2) Text Extraction from Images Captured via Mobile and Digital Devices by Jian Yuan, Yi Zhang, Kok Kiong Tan, Tong Heng Lee. et.all.

In this paper, the application will only recognize a few commonly used non-italic font types which are usually the case in natural scene images.

Disadvantages:
The space for improvement on robustness against font types and font thickness, as well as translations as sentences instead of each word using machine translation techniques.

3) A Robust Algorithm for Text Extraction from Images by Najwa-Maria Chidiac, Pascal Damien, Charles Yaacoub. et.all. A Robust algorithm that detects text from natural scene images and extracts them regardless of the orientation is proposed.

Disadvantages:
The algorithm failed in detecting text with shadowing effect, as well as characters with very small size and/or thin strokes.

4) Automatic detection and translation of text from natural scenes by Jie Yang, Xilin Chen, Jing Zhang, Ying Zhang, Alex Waibel. Et.all. An automatic sign translation system utilizes a camera to capture the image with signs, detects signs in the image, recognizes signs, and translates results of sign recognition into a target language. Such a system relies on technologies of sign detection, OCR, and machine translation.

Disadvantages:
The confidence of the sign detection can be improved by incorporating the OCR engine in an early stage.

5) An Adaptive Machine Translator for Multilingual Communication by Ryan Lane , Ajay Bansal.

Building a machine translator generator for multilingual communication, i.e. developing a system whose inputs are linguistic descriptions of a desired source and target language and whose output is a program that translates between the two natural languages.

Disadvantages:
The system were to be improved beyond a basic translator between relatively small subsets of two languages in order to more rigorously explore the domain of machine translation—the original research goal.

IV. PROPOSED WORK

a. Overview

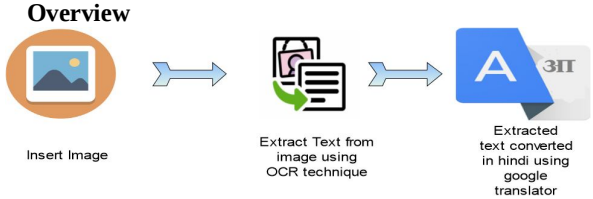

Fig.1: Architecture Diagram

To carry out this task we have four modules those are Image Capture, Text Identification, Language conversion, PDF generation. A mobile camera is used to capture the image. It is important to learn how to use a mobile camera properly so that you can convert a text to image effectively and get the most accurate results. The text is given as input and image is get as output the input image is first pre-processed to remove the noise present in the image. The image is converted into a gray scale image which can then be converted into binary image. Tesseract is an Optical Character Recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006.Tesseract is considered one of the most accurate open source OCR engines currently available. The total count of support languages to over 60.It is the tool used to extract the text from an image. After extracting the words from image by using the Optical Character Recognition (OCR) Engine, those words are translated into known language to do this the Bing Translator API service is used. This is a free service. It provides many libraries for translation. The first thing to remember is that translation is the transfer of meaning from one language to another.

V. ALGORITHM

1) OCR

Documents of our interest are typically electronic images of paper documents obtained with a scanner. A document is associated with a schema, as follows. The schema describes the information to be extracted from the document and consists of a set of typed elements e: for each element, the document contains zero or one value v. For example, a schema could be date, total Amount, document Number, respectively with types date, currency and number; a document with this schema could contain the values "7/2/2011", "23,79" and no value for the respective elements. Executing an OCR procedure on the electronic image of a document we obtain a set of strings {l1, l2. . . ln}.

For each element e of the schema, we associate the candidate string l to that element. The description of the system that automatically associates each element e with the candidate string l is beyond the scope of this paper. For each searched element, it may be l 6= v, because of the following reasons:

• l may contain v and extra-text that is not part of v.

For example l = "date:21/12/2008" while v = "21/12/2008".

• l may not contain v due to OCR errors. These errors can be of two types: – segmentation error: different line, word or character spacing's lead to misrecognitions of white-spaces, causing segmentation errors

(e.g., l = "076 4352 056 C" while v = "0764352056C").

Misrecognition of characters: low print quality, dirt and font variations prevent an accurate recognition of characters (e.g., l

= "|9,5SG" while v = "19,556" or l = "IOAS/0B127" while v = "105/08127"). While the segmentation and misrecognition problem may occur only with digitized documents, the extra-text problem may occur also with digitally born documents. We propose a solution that uses a suite of syntactic and semantic checks in order to detect and correct these OCR generated errors. Our system is designed to be modular and extensible, so as to make it possible to augment and improve the domain knowledge encoded in the module as well as to accommodate further application-specific rules beyond those currently embedded in the system. A high-level description of this step follows, full details are provided in the next sections

For each element, we generate a set of values {v ∗ 1 , v∗ 2 , . . . , v∗ n} that, due to OCR errors, might have led to the extraction of l. This set is generated by applying to the extracted string l a number of possible substitutions as encoded in a predefined table of substitution rules. This table encodes a domain-knowledge about possible misrecognitions of characters. Then, we perform a suite of syntactic and semantic checks to exclude from the previous set all those values that do not satisfy at least one of these checks. We denote by V ∗ the resulting set of candidate values. These syntactic and semantic checks are encoded in Boolean functions tailored to the type of the element to be extracted. We have implemented these functions for the following elements: data, number, vat Number, currency, fiscal code (a unique id assigned to each person that lives in Italy, whose structure follows certain constraints)

Optical Character Recognition (OCR) is a technique used to interpret scanned documents into computer readable text. This thesis focuses on using OCR to interpret invoices and their content. The OCR-process had to result in the invoice structure to be relatively alike how it was structured before the process. It was of uttermost importance that the structure of the invoice did not differ after the OCR process.

Steps in OCR:

1. Loading any image format (bmp, jpg, png) from given source. Then convert the image to grayscale and binarize it using the threshold value (Otsu algorithm). //completed(How to remove noise from output Image???)

2. Detecting image features like resolution and inversion. So that we can finally convert it to a straightened image for further processing. (completed the code of rotation of Image but not able to detect Image angle about which we have to rotate the Image,So still working on angle detection part)
3. Lines detection and removing. This step is required to improve page layout analysis, to achieve better recognition quality for underlined text, to detect tables, etc.(Decided To Complete that part in End)
4. Page layout analysis. In this step I am trying to identify the text zones present in the image. So that only that portion is used for recognition and rest of the region is left out.
5. Detection of text lines and words. Here we also need to take care of different font sizes and small spaces between words.
6. Recognition of characters. This is the main algorithm of OCR; an image of every character must be converted to appropriate character code. Sometimes this algorithm produces several character codes for uncertain images. For instance, recognition of the image of "I" character can produce "I", "|" "1", "l" codes and the final character code will be selected later.
7. Saving results and convert to language and so on.

Methodologies

This application contains three steps.

1. Take a photo image of the unknown language text which you want to translate(either handwritten or printed material),
2. Tesseract is an open source Optical Character Recognition (OCR) technology, which is used to extract the text from the image then Google API and Bing API is used for translation of language.
3. The translated text is generated in text format.

Tools and Technologies Used:

This application is mounted on the Internet, to user has to make sure that the machine, which he is using, is connected to Internet through Lease Line, Telephone line or Cable.

Also, Microsoft Internet Explorer 4.0 and above or Netscape Navigator 4.74 and above must be installed on the machine.

Mathematical Model:

TF-IDF is a way of scoring the vocabulary so as to provide adequate weight to a word in proportion of the impact it has on the meaning of a sentence. The score is a product of 2 independent scores, term frequency(tf) and inverse document frequency (idf)

TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Picture Credit- Google

Term Frequency (TF): Term frequency is defined as frequency of word in the current document.

Inverse Document Frequency (IDF): is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is calculated as log (N/d) where, N is total number of documents and d is the number of documents in which the word appears.

VI. CONCLUSION

This is the discussion about optical character recognition techniques to translate the text from unknown language text into known language. The system has the capability to recognize characters with accuracy exceeding 90% mark. The advantage of this system is that it is easily portable and its scalability which can recognize various languages and also help in translating the text in different languages. The accurate recognition is directly depending on the nature of the material to be read and by its quality.

VII. FUTURE SCOPE

In Future we make translation in many local languages. Translated text is in play as a voice message.

VIII. REFERENCE

[1]. J. Liang, D. Doermann, H. Li: Camera-based analysis of text and documents: A survey. International Journal on Document Analysis and Recognition 7(2005):84-104.
[2]. Epshtein, B., Ofek, E., Wexler, Y. Detecting text in natural scenes with stroke width transform. CVPR 2010.
[3]. Y.-F. Pan, X. Hou, and C.-L. Liu. Text localization in natural scene images based on conditional random field. ICDAR 2009.
[4]. M. Koga, R. Mine, T. Kameyama, T. Takahashi, M. Yamazaki, T. Yamaguchi: Camera-based Kanji OCR for mobile-phones: practical issues. ICDAR 2005.
[5]. X. Chen, J. Yang, J. Zhang, and A. Waibel. Automatic detection and recognition of signs from natural scenes. IEEE Transactions on Image Processing, 13(1):87–99, 2004.
[6]. K. Jung, K. I. Kim and A. K. Jain. Text information extraction in images and video: a survey. Pattern Recognition, 37(5): 977-997, 2004.
[7]. D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2(60):91–110, 2004.
[8]. K. Mikolajczyk, and C. Schmid, A Performance Evaluation of Local Descriptors, In CVPR, 2003.
[9]. Ohya, J., Shio, A., and Akamatsu, A., Recognition of characters in scene images. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 2, pp. 214- 220,1994.
[10]. Sato, T., Kanade, T., Hughes, E.K., and Smith, M.A., Video OCR for digital news archives. IEEE Int. Workshop on Content-Based Access of Image and Video Database, 1998.