APPROACH TO TEXT EXTRACTION FROM IMAGE

Disha Bhat  
School of Computing and Information Technology  
REVA UNIVERSITY, Bangalore, India

Charitha D  
School of Computing and Information Technology  
REVA UNIVERSITY, Bangalore, India

Dimple M K  
School of Computing and Information Technology  
REVA UNIVERSITY, Bangalore, India

Amruthashree R V  
School of Computing and Information Technology  
REVA UNIVERSITY, Bangalore, India

Shruthi G  
School of Computing and Information Technology  
REVA UNIVERSITY, Bangalore, India

Abstract: The multimedia resources in a database and on the web are increasing. The multimedia resources can be images and videos. It has become a very difficult task to develop the effective methods to manage as well as to retrieve these resources by their content. The Text which is an important object which carries high-level semantic information which is useful for this task. The current technology is optical character recognition (OCR) is used to convert machine generated text which is printed against clean background to computer readable form (ASCII). But, text is often printed against shaded or textured backgrounds or is embedded in images. Examples include maps, photographs, advertisements, videos etc. Current document segmentation and recognition technologies cannot handle these situations well. Our system takes advantages of the distinctive characteristics of text that make it stand out from other image material that is, text possesses certain frequency and orientation information. We will first clean the image by changing the contrast and gradient of the image. Now the objects in the images are identified and numbered. Further in the text recognition process, these numbered objects are segregated into text and non-text. Later the recognised text is reconstructed to form a meaningful text present in the image. Also we are focusing on extracting the text such that certain portion of the images such as logos etc is retained. This is done by calculating the pixels of the required portion of the image to be retained and then training the system in such a way that it extracts all the text except the portion of the image to be retained.

Keywords: image, text, extraction

1. INTRODUCTION

Today, the vast majority of the data is accessible on paper or as photos or recordings. Expansive data is put away in pictures. The present innovation is confined to removing content against clean foundations. Subsequently, there is a requirement for a framework to extricate content from general foundations. These are various applications in which text extraction is useful. These applications include digital libraries, multimedia systems, information retrieval system and geographical information system. The role of text that can be directly highlighted to user or fed into an optical character reader module for recognition. In this a new system is proposed which extracts text in images and also retains certain portion of the image such as a logo. The system takes coloured images as input. It detects text on the basis of certain text features: text possesses certain frequency and orientation information. Text shows spatial cohesion characters of the same text string are of similar heights, orientation and spacing. The image is then cleaned up so that the text stands out, to retain the logo or a certain part of the image. First calculate the pixels of the part of the image to be retained and then instruct the system in such a way that it extracts all other text from the image except that certain part. The optical character recognition technology has three types. The OCR, which targets on typewritten text, one glyph or character at a time. The ICR (Intelligent character recognition) which targets handwritten print script or cursive text one glyph or character at a time. The IWR (Intelligent Word Recognition) which targets one word at a time. All these technologies are unable to properly reconstruct the text after text recognition. So this motivated us to build such a system which can reconstruct the text properly.

- To improve the current and existing technology of OCR.
- To make the text reconstruction process more reliable and accurate.
- To retain certain part of the image such as logo this need not be extracted.
• To make a cost effective software and make it useful for people.

II. LITERATURE SURVEY

The overwhelming volume of paper based data in corporation and offices challenges their ability to manage documents and records. Computers work efficiently and faster than human operators. Computers understand alpha numeric characters therefore the English letter should be into computer recognisable form. OCR allows us to convert a document into electronic text, it effectively separates individual characters. OCR has become one of the most successful application in pattern recognition and AI.

Figure 1: OCR Generation

1st Generation OCR System
- Generation was IBM 1418
- Designed to read a special IBM font 407
- Recognition method was template matching

2nd generation OCR System
- We were able to recognise regular machine printed and hand printed characters
- Characters and numbers were limited

3rd Generation OCR System
- Large documents of handwritten were recognised easily
- Low cost and high performance

4th Generation OCR System (OCR’s today)
Documents can be intermediated with text, graphics, tables and mathematical symbols, colour documents and noisy documents.

Figure 2: Tasks involved in OCR

Image acquisition: Input image for OCR system might be acquired by scanning document or by capturing photograph of document. This is also known as digitization process.

Pre-processing: Pre-processing comprise arrangement of tasks and it used to upgrade a picture and make it reasonable for division. Commotion gets presented amid archive age. So Proper channel like mean channel, min-max channel, Gaussian channel and so forth might be connected to expel clamor from archive. Binarization process changes over dark scale or hued picture to high contrast picture. To upgrade perceivability and basic data of character Binary morphological activities like opening, shutting, diminishing, gap filling and so forth might be connected on picture. In the event that examined picture isn’t be flawlessly adjusted, so we have to adjust it by performing incline edge amendment. Info report might be resized in the event that it is too extensive in size to lessen measurements to enhance speed of preparing.

Segmentation: Character segmentation performs an operation of decomposition of an image into Sub images of individual symbols. It is one of the decision processes in a system for optical character recognition (OCR). Its decision that a pattern isolated from the image is that of a character or some other identifiable unit. Generally document is processed in hierarchical way. At first level lines are segmented using row histogram. From each row, words are extracted using column histogram and finally characters are extracted from words. Accuracy of final result is highly depends on accuracy of segmentation.

Feature extraction: Feature extraction is the important part of any pattern recognition application. Feature extraction techniques like Linear Discriminant Analysis (LDA), Principle Component Analysis (PCA),Independent Component Analysis (ICA), Chain Code (CC), Scale
Invariant Feature Extraction (SIFT), Gradient based features, Histogram might be applied to extract the features of individual characters. These features are used to train the system.

**Classification:** When image is provided as input to OCR system, its features are extracted and given as an input to the trained classifier like artificial neural network or support vector machine. Classifiers compare the input feature with stored pattern and find out the best matching class for input.

**Post processing:** This progression isn't necessary; it enhances the precision of acknowledgment. Sentence structure examination, semantic investigation sort of more elevated amount ideas may be connected to check the setting of perceived character.

**OCR Classification based on Fonts**
Based on the OCR systems’ capability to recognize different character sets, a classification [Line, 1993], by the order of difficulty is as follows.

- **Fixed font OCRs** OCR machines of this category deal with the recognition of characters in only one specific typewritten font. Examples of such fonts are OCR-A, OCR-B, Pica, Elite, etc. These fonts are characterized by fixed spacing between each character. The OCR-A and OCR-B are the American and European standard fonts specially designed for optical character recognition, where each character has a unique shape to avoid ambiguity with other characters similar in shape. Using these character sets, it is quite common for commercial OCR machines to achieve a recognition rate as high as 99.99% with a high reading speed. The first generation OCRs were fixed font machines, and the methods applied were usually based on template matching and correlation.

- **Multi-font OCRs** Multi-text style OCR machines perceive characters from in excess of one textual style, rather than a settled text style framework, which could just perceive images of one particular textual style. For the prior age OCRs, the breaking point in the quantity of perceived text styles was because of the example acknowledgment calculation utilized: format coordinating, which required that a library of bit outline of each character from every textual style was put away (prerequisite of a tremendous database). The exactness is very great, even on corrupted pictures, as long as the text styles in the library are chosen with mind.

- **Omnifont OCRs** An omni text style OCR machine can perceive images of most non-adapted textual styles without maintaining gigantic databases of particular textual style data. Generally omni text style innovation is described by the utilization of highlight extraction. The database of an omni text style framework will contain a depiction of every image class rather than the images themselves. This gives adaptability in programmed acknowledgment of characters from an assortment of text styles. Various current OCR-frameworks for English claim to be omni text style. In spite of the fact that omni text style is the normal term utilized for these OCRs, this does not imply that they perceive characters from every single existing textual style.

**Textual style characterization can diminish the quantity of elective shapes for each class, driving basically to single-text style character acknowledgment [Zhu, Tan and Wang, 2001]. Following is the review of methodologies utilized as a part of the writing for perceiving text styles in English.**

**III. SYSTEM ANALYSIS AND DESIGN**
Optical character acknowledgment (OCR) strategy has been utilized as a part of changing over printed content into editable content. OCR is exceptionally valuable and prominent strategy in different applications. Exactness of OCR can be subject to content pre-preparing and division calculations. Here and there it is hard to recover content from the picture due to various size, style, introduction, complex foundation of picture and so forth. We start this paper with a presentation of Optical Character Recognition (OCR) strategy, History of Open Source OCR device Tesseract, design of it and test consequence of OCR performed by Tesseract on various types pictures are talked about. We finish up this paper by similar investigation of this device with other business OCR apparatus Transym OCR by considering vehicle number plate as info. From vehicle number plate we endeavoured to extricate vehicle number by utilizing Tesseract and Transym and thought about these devices in light of different parameters. Optical Character Recognition (OCR), Open Source, DLL, Tesseract, Transym Optical character Recognition (OCR) is a change of checked or printed content pictures, written by hand message into editable content for additionally handling. This innovation enables machine to perceive the content consequently. It resembles mix of eye and psyche of human body.

**APPLICATION**
Optical character recognition has been applied to a number of applications. Some of them have been explained below.

An eye can see the content from the pictures however really the mind forms and in addition deciphers that separated content read by eye. Being developed of automated OCR framework, couple of issues can happen. In the first place: there is almost no unmistakable distinction between a few letters and digits for PCs to get it. For instance it may be troublesome for the PC to separate between digit "0" and letter "o". Second: It may be exceptionally hard to separate content, which is inserted in extremely dull foundation or imprinted on different words or illustrations. In 1955, the principal business framework was introduced at the per user's process, which utilized OCR to include deals report into a PC and after that after OCR strategy has turned out to be exceptionally useful in modernizing the physical office archives.
Table 1: Comparison of Optical Character Recognition Software's

<table>
<thead>
<tr>
<th>Name</th>
<th>Founded year</th>
<th>Online</th>
<th>Compatibility</th>
<th>Programming Language</th>
<th>SDK?</th>
<th>Fonts</th>
<th>Output formats</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pytesseract</td>
<td>1985</td>
<td>No</td>
<td>Windows MacOS X, Linux BSD</td>
<td>C++, C</td>
<td>Yes</td>
<td>Any printing font</td>
<td>Text in OCR, PDF, others with different user interface or the API</td>
<td>100+</td>
</tr>
<tr>
<td>Screenworm</td>
<td>2013</td>
<td>No</td>
<td>MacOS X</td>
<td>Objective C++</td>
<td>No</td>
<td>?</td>
<td>TXT</td>
<td>57</td>
</tr>
</tbody>
</table>

- **Healthcare**
  Medicinal services experts dependably need to manage extensive volumes of structures for every patient, including protection frames and also broad wellbeing shapes. To stay aware of the majority of this data, it is helpful to include significant information into an electronic database that can be gotten to as vital. Frame handling apparatuses, fueled by OCR, can remove data from structures and place it into databases, with the goal that each patient's information is quickly recorded.

- **Optical Music Recognition**
  Initially it was aimed towards recognizing printed sheets which can be edited into playable form with the help of electronic methods. It has many applications like processing of different classes of music, large scale digitization of musical data and also it can be used for diversity in musical notation.

- **Legal Industry**
  OCR is utilized as a part of Legal industry for digitize records, and specifically entered to PC database. Lawful experts can additionally look reports required from enormous databases by just composing a couple of catchphrases.

- **Automatic Number Recognition**
  Automatic number plate recognition is used as a technique making use of optical character recognition on images to identify vehicle registration plates. They are used by various police forces and as a method of electronic toll collection on pay-per-use roads and cataloging the movements of traffic or individuals

- **Handwriting Recognition**
  It is the ability of a computer system which scans the image of handwritten text by scanner and extracts only handwritten character from that image.

IV. RESULTS

![Figure 3: input image](image1)

![Figure 4: Output image of Figure 3](image2)

![Figure 5: Input image](image3)
IV. CONCLUSION

This work goes for takes points of interest of the particular attributes of content that influence it to emerge from other picture material that is, content has certain recurrence and introduction data. We will first clean the image by changing the contrast and gradient of the image. Now the objects in the images are identified and numbered. Further in the text recognition process, these numbered objects are segregated into text and non-text. Later the recognised text is reconstructed to form a meaningful text present in the image. Also we are focusing on extracting the text such that certain portion of the images such as logos etc is retained. This is done by calculating the pixels of the required portion of the image to be retained and then training the system in such a way that it extracts all the text except the portion of the image to be retained obtain average accuracy in the work. Although the results of OCR System are not good, they are not bad either, indicating that the OCR Technique is not awed. More training data may improve robustness and accuracy.

V. REFERENCES: