RESEARCH ARTICLE OPEN ACCESS

Text Recognition from an Image

Shrinath Janvalkar, Paresh Manjrekar, Sarvesh Pawar, Prof. Laxman Naik
Department Of Computer Engineering[1, 2, 3, 4]
RMCET(Mumbai University)
Ambav, Devrukh, India[1, 2, 3, 4]

ABSTRACT
To achieve high speed in data processing it is necessary to convert the analog data into digital data. Storage of hard copy of any document occupies large space and retrieving of information from that document is time consuming. Optical character recognition system is an effective way in recognition of printed character. It provides an easy way to recognize and convert the printed text on image into the editable text. It also increases the speed of data retrieval from the image. The image which contains characters can be scanned through scanner and then recognition engine of the OCR system interpret the images and convert images of printed characters into machine-readable characters [8].It improving the interface between man and machine in many applications.

I. INTRODUCTION

Character recognition is one of the most interesting areas of pattern recognition and artificial intelligence. Optical Character Recognition extracts the relevant information and automatically enters it into electronic database instead of the conventional way of manually retyping the text. Optical Character Recognition is a vast field with a number of varied applications such as invoice imaging, legal industry, banking, health care industry etc. OCR is also widely used in many other fields like Captcha, Institutional repositories and digital libraries, Optical Music Recognition without any human correction or human effort, Automatic number plate recognition and Handwritten Recognition [6]. It contributes immensely to the advancement of an automation process and can improve the interface between man and machine in numerous applications. Several research works have been focusing on new techniques and methods that would reduce the processing time while providing higher recognition accuracy. Now it is possible to scan documents as an image and to make it editable and searchable for further information processing.

II. OBJECTIVE

The objective of OCR software is to recognize the text and then convert it to editable form. Thus, developing computer algorithms to identify the
character in the text is the principal task of OCR. A document is first scanned by an optical scanner, which produces an image form of it that is not editable. Optical character recognition involves. Translation of this text image into editable character codes such as ASCII [4]. The below diagram shows the processing mechanism of OCR system (Fig. 1).

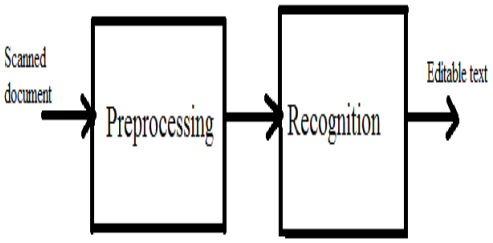


Fig. 1. OCR Engine

III. LITERARURE SURVEY

3.1 First generation OCR systems

The first commercialized OCR of this generation was IBM 1418, which was designed to read a special IBM font407. The recognition method was template matching, which compares the character image with a library of prototype images for each character of each font [5].

3.2 Second generation OCR systems

Next generation machines were able to recognize regular machine-printed and hand printed characters. The character set was limited to numerals and a few letters and symbols. Such machines appeared in the middle of 1960s to early 1970s [5].

3.3 Third generation OCR systems

For the third generation of OCR systems, the challenges were documents of poor quality and large printed and hand-written character sets. Low cost and high performance were also important objectives. Commercial OCR systems with such capabilities appeared during the decade 1975 to 1985[5].

3.4 OCRs Today (Fourth generation OCR systems)
The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, tables and mathematical symbols, unconstrained handwritten characters, color documents, low-quality noisy documents, etc. Among the commercial products, postal address readers, and reading aids for the blind are available in the market [5].

IV. TASK INVOLVED IN OCR

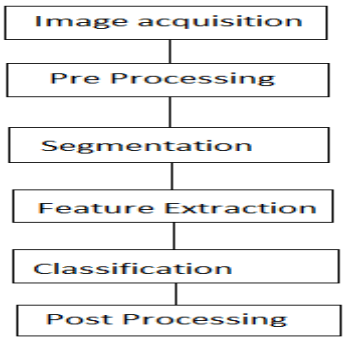


Fig. 2. OCR Processing

The above figure shows different processes which are done in OCR system (Fig. 2).

4.1 Image acquisition
Input image for OCR system might be acquire by scanning document or by capturing photograph of document. This is also known as digitization process [11].

4.2 Preprocessing
Preprocessing consist series of operations and it used to enhance an image and make it suitable for segmentation. Noise get introduced during document generation. So Proper filter like mean filter, min-max filter, Gaussian filter etc. may be applied to remove noise from document.Binarization process converts gray scale or colored image to black and white image. To enhance visibility and structural information of character Binary morphological operations like opening, closing, thinning, hole filling etc. may be applied on image. If scanned image is not be perfectly aligned, so we need to align it by performing slant angle correction. Input document may be resized if it is too large in size to reduce dimensions to improve speed of processing [11].

4.3 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 [11].

4.4 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 [11].

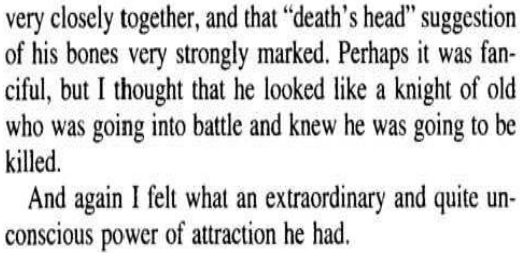
4.5 Classification

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 [11].

4.6 Post processing
This step is not compulsory; it helps to improve the accuracy of recognition. Syntax analysis, semantic analysis kind of higher level concepts might be applied to check the context of recognized character [11].

V. RESULT: SECTION OF OCR IMAGE

Input as Image:


very closely together, and that "death's head" suggestion of his bones very strongly marked. Perhaps it was fanciful, but I thought that he looked like a knight of old who was going into battle and knew he was going to be killed.

And again I felt what an extraordinary and quite unconscious power of attraction he had.

Fig. 3. Input to OCR System

Result:

[1] "veo etoyels Ioke oer, end net r`deao,k head fg suhgestion" very closely together, and that ``deaths's head " suggestion
[2] "oI gtd genep eaw stsougly markod. serhass lI was scnw" of his bones very strongly marked. Perhaps it was fan-
[3] "cifol, hmt I thoOphi ihat ha looser lthe a knight of old" ciful, but I thought that he looked like a knight of old

[4] "wk, was goine into batIta anr snew he Wak geing so he" who was going into battle and knew he was going to be
[5] "... apxhn I tcit what an enH aorhi.Mwy ans suiie unm" ... again, I felt what an extraordinary and quite un-
[6] "eoaserouk poWer nf attracigHn he had." conscious power of attraction he had.

VI. APPLICATIONS

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

6.1 *Legal Industry*

OCR is used in Legal industry for digitize documents, and directly entered to computer database. Legal professionals can further search documents required from huge databases by simply typing a few keywords [6].

6.2 Healthcare

Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR, are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded [6].

6.3 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 [6].

6.4 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 [6].

6.5 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 [7].

VII. CONCLUTION

Although results of OCR System are not good, they are not that bad either, indicating that the OCR technique is not awed. More training data may improve robustness and accuracy.

REFFERENCES

[1] Optical Character Recognition using Neural Networks Deepayan Sarkar University of Wisconsin MadisonECE 539 Project, Fall 2003.
[2] "Evaluation of OCR Algorithms for Images with Different Spatial Resolutions and Noises" School of Information Technology and Engineering Faculty of Engineering University of Ottawa©Qing Chen, Ottawa, Canada, 2003.
[3] "A Neural Network Implementation of Optical Character Recognition" Technical Report Number CSSE10-05 COMP 6600 – Artificial Intelligence Spring 2009.
[4] "Optical Character Recognition Techniques: A Survey" Sukhpreet Singh M.tech Student, Dept. of Computer Engineering, YCOE Talwandi Sabo BP. India.
[5] OCR System: A Literature Survey
[6] "Survey of OCR Applications" by Amarjot Singh, ketan bacchuwar, Akshay bhasin.
[7] M.D. Ganis, C.L. Wilson, J.L. Blue, "Neural network-based systems for handprint OCR applications" in IEEE Transactions on Image Processing, 1998, Vol: 7, Issue: 8, p.p. 1097 – 1112.
[8] "Performance Characterization and Acceleration of Optical Character Recognition on Handheld Platforms" Sadagopan Srinivasan, Li Zhao, Lin Sun, Zhen Fang, Peng Li, Tao Wang,Ravishankar Iyer, Ramesh Illikkal, Dong LiuIntel Corporation.
[9] "Implementing Optical Character Recognition on the Android Operating System for Business Cards" Sonia Bhaskar, Nicholas Lavassar, Scott GreenEE 368 Digital Image Processing.
[10] "A Comparative analysis of feature extraction techniques for handwritten character recognition" Rajbala Tokas1, Aruna Bhadu2 M.Tech*(CS), Swami Keshwanand Institute of Technology, Jaipur, Rajasthan, India, M.Tech*(SE) Govt. Engineering College.
[11] "A Literature Review on Hand Written Character Recognition" by Mansi shah & Gordhan B Jethava Department of Computer Science & Engineering Parul Institute of Technology, Gujarat, India. Information Technology Department Parul Institute of Engg. & Technology, Gujarat, India.