ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Optical Character Recognition

Shalin A. Chopra[1], Amit A. Ghadge[2], Onkar A. Padwal[3], Karan S. Punjabi[4], Prof. Gandhali S. Gurjar[5]

Student, Dept. Of Computer Engineering, Sinhgad Academy of Engineering, Pune, India [1]
Student, Dept. Of Computer Engineering, Sinhgad Academy of Engineering, Pune, India [2]
Student, Dept. Of Computer Engineering, Sinhgad Academy of Engineering, Pune, India [3]
Student, Dept. Of Computer Engineering, Sinhgad Academy of Engineering, Pune, India [4]
Assistant Professor, Dept. Of Computer Engineering, Sinhgad Academy of Engineering, Pune, India [5]

Abstract: At the present time, keyboarding remains the most common way of inputting data into computers. This is probably the most time consuming and labour intensive operation. OCR is the machine replication of human reading and has been the subject of intensive research for more than three decades. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text. It is a method of digitizing printed texts so that they can be electronically searched and used in machine processes. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining. This paper presents a simple, efficient, and less costly approach to construct OCR for reading any document that has fix font size and style or handwritten style. To achieve efficiency and less computational cost, OCR in this paper uses database to recognize English characters which makes this OCR very simple to manage.

Keywords: Scanned images, digitizing, translation, machine –encoded text, fix font, handwritten style, text-to-speech.

I. INTRODUCTION

The advancements in pattern recognition has accelerated recently due to the many emerging applications which are not only challenging, but also computationally more demanding, such evident in Optical Character Recognition (OCR), Document Classification, Computer Vision, Data Mining, Shape Recognition, and Biometric Authentication, for instance.[13] The area of OCR is becoming an integral part of document scanners, and is used in many applications such as postal processing, script recognition, banking, security (i.e. passport authentication) and language identification. The research in this area has been ongoing for over half a century and the outcomes have been astounding with successful recognition rates for printed characters exceeding 99%, with significant improvements in performance for handwritten cursive character recognition where recognition rates have exceeded the 90% mark. [13]

Nowadays, many organizations are depending on OCR systems to eliminate the human interactions for better performance and efficiency. Optical Character Recognition also referred to as OCR is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the document [2]. Documents are scanned using a scanner and are given to the OCR systems which recognizes the characters in the scanned documents and converts them into ASCII data. [2]

In OCR a database is used at the back end for recognition. In proposed system the process consists of following processing steps: (1) Scanning of Image, (2) Pre-Processing of Image (3) Character Extraction (4) Feature Extraction and Recognition (5) Post-Processing. [1]

In the document scanning step, a scanner is used to scan the handwritten or printed documents. The quality of the scanned document depends up on the scanner. So, a scanner with high speed and color quality is desirable. The recognizing process includes several complex algorithms and previously loaded templates and dictionary which are crosschecked with the characters in the document and the corresponding machine editable ASCII characters. The verifying is done either randomly or chronologically by human Intervention [2].

Optical Character Recognition is classified into two types, Offline recognition and Online recognition. In offline recognition the source is either an image or a scanned form of the document whereas in Online recognition the successive points are represented as a function of time and the order of strokes are also available [11]. Here in this paper only offline recognition is dealt.

The proposed OCR system provides the following features: [12]

- No more retyping,
- Quick Digital Searches,
- Edit Text,
- Save Space.

II. CONSTRUCTION OF OCR SYSTEM

A. Pre Processing

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image (Binarization). Practically any scanner is not perfect; the scanned image may have some noise. This noise may be due to some unnecessary details present in the image. By applying suitable methods the denoised image is produced. The denoised image thus obtained is saved for further processing [8].

B. Character Extraction

The pre-processed image serves as the input to this and each single character in the image is found out [1].

C. Recognition

The image from the extraction stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image. [1]

D. Post Processing

After the recognition stage, if there are some unrecognised characters found, those characters are given their meaning in the post-processing stage. Extra templates can be added to the system for providing a wide range of compatibility checking in the systems database [2].

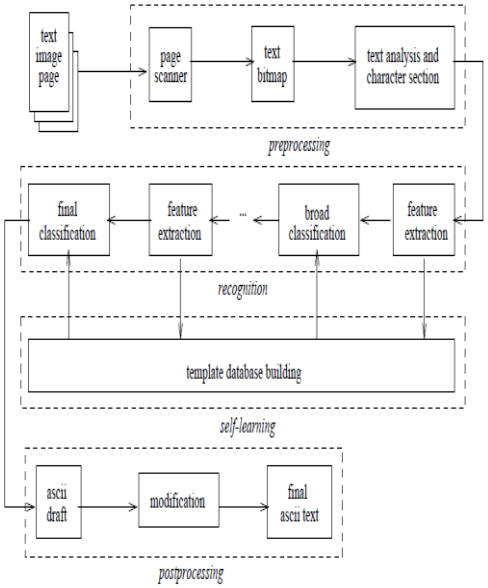

Fig 1. Construction of OCR system

III. DATA FLOW DIAGRAMS

The DFD serves two purposes:

1. To provide an indication of how data are transformed as they move through the system.
2. To depict the function and sub-functions that transforms the data.

They serve as basis for the functional as well as information flow modelling.

Fig 2. Level 0 DFD

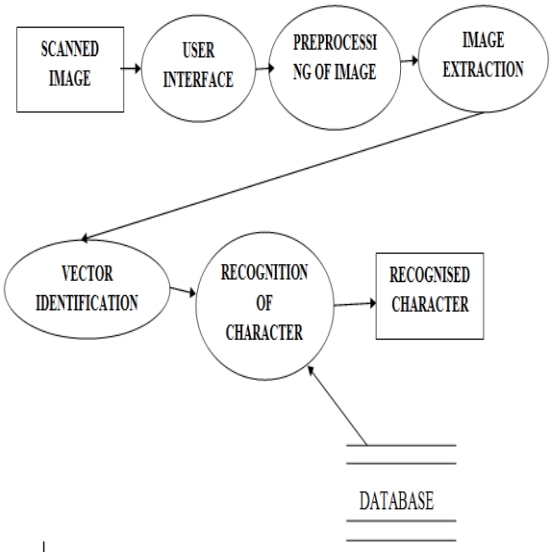

Fig 3. Level 1 DFD

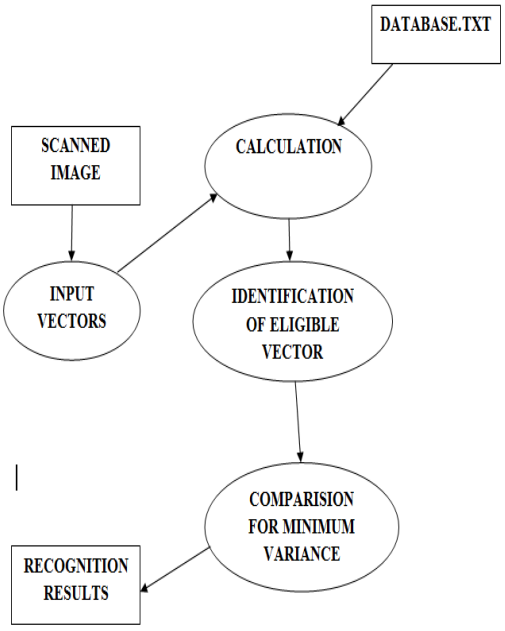

Fig 4. Level 2 DFD

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

IV. CONCLUSION

This paper tells about OCR system for offline handwritten character recognition. The systems have the ability to yield excellent results. Preprocessing techniques used in document images as an initial step in character recognition systems were presented. The feature extraction step of optical character recognition is the most important. It can be used with existing OCR methods, especially for English text. This system offers an upper edge by having an advantage i.e. its scalability, i.e. although it is configured to read a predefined set of document formats, currently English documents, it can be configured to recognize new types.

Future research aims at new applications such as online character recognition used in mobile devices, extraction of text from video images, extraction of information from security documents and processing of historical documents.

ACKNOWLEDGMENT

The authors would like to thank Department of Computer Engineering, SAE and indebted to our guide Prof. Gandhali. S. Gurjar for her guidance and sagacity without which this paper would not have been designed. She provided us with valuable advice which helped us to accomplish the design of this paper. We are also thankful to our HOD Prof. B. B. Gite (Department of Computer Engineering) for his constant encouragement and moral support. Also we would like to appreciate the support and encouragement of our colleagues who helped us in correcting our mistakes and proceeding further to produce the paper with the required standards.

REFERENCES

[1] "α-Soft: An English Language OCR", 2010 Second International Conference on Computer Engineering and Applications. Junaid Tariq, Umar Nauman Muhammad Umair Naru.

[2] "A Review on the Various Techniques used for Optical Character Recognition", Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi/ International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.

[3] "Character Recognition in practice Today and Tomorrow", 1996, Udo Miletzki, Siemens Electrocom GmbH D-78767 Konstanz, Germany.

[4] "Prototype Extraction and Adaptive OCR" IEEE Transaction on pattern analysis and Machine Intelligence, VOL. 21, NO. 12, DECEMBER 1999, Yihong XU, Member, IEEE, George Nagy, Senior Member, IEEE.

[5] "Contextual Focus for Improved Recognition of Hand-Filled Forms", 1999. Wing Seong Wong, Nasser Sherkat, Tony Allen IRIS, Department of Computing.

[6] "Image processing Algorithms for Improved Character Recognition and Components Inspection", 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), Anima Majumder.

[7] "A System for Automated Data Entry from Forms", 1996 IEEE Proceedings of ICPR '96, Raymond A. Lorie, V. P. Riyaz , Thomas K. Truong.

[8] "Combination of Document Image Binarization Techniques", 2011 International Conference on Document Analysis and Recognition.

[9] ICAR: Identity Card Automatic Reader, 200 1 IEEE, Josep Lladbs, Felipe Lumbreras, Vicente Chapaprieta, Joan Queralt.

[10] "Implementing Optical Character Recognition on the Android Operating System for Business Cards", IEEE 2010, Sonia Bhaskar, Nicholas Lavassar, Scott Green EE 368 Digital Image Processing.

[11] "Document Analysis and Recognition", 2005. Eighth International Conference on 29 Aug.-1 Sept. 2005, Alon, Jonathan.

[12] en.wikipedia.org/wiki/Optical_character_recognition

[13] Pre-processing Techniques in Character Recognition, Yaseer Alginahi.