**Image to Text Conversion Using OCR Technique**

**Abstract**

To detect text characters from an image is very crucial task. One procedure using OCR is to detect text characters from a document image. However, some documents might come with embedded background images which often mislead the algorithms of character detection. For example, small dots or sharp edges from the background image are often bound boxed as characters and passed to the next stage of the OCR pipeline, which causes an error chain. Motivated by this observation, we present a novel and cost-effective image preprocessing method to accomplish the task. We will first use the document images and preprocess this. That is we will convert color images to gray and threshold it. This way, background images can be removed effectively without losing the quality of text characters. Using OCR will successfully extract all the characters from image. This concept of extracting the contents of document image and display that in text area.

**Keyword:** Document image, OCR, image preprocessing, gray image, threshold image;

**Introduction**

Now-a-days, there is growing demand for the software systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. These days there is a huge demand in “storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process”. One simple way to store information in these paper documents in to computer system is to first scan the documents. Whenever we scan the documents through the scanner, the documents are stored as images format in the computer system. These images containing text cannot be edited by the user. But to reuse this information it is very difficult for computer system to read the individual contents and searching the contents form these documents line-by-line and word-by-word**.**

OCR technology has a broad range of applications in document processing. Many document images are embedded with background images, e.g., checks, deposit books, drive licenses, passports, certificates, etc. While the background image enhances the document's security or visual effects, it causes difficulties for OCR applications. Some parts of the background image could be bound-boxed as characters, which leads to immediate wrong recognition and causes troubles in the following processing steps of the OCR pipeline. Therefore, it is very important to preprocess the documents by removing the background images before text detection. There are several existing methods on text extraction from complex background.

**Literature Survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No** | **Author Name/ Paper Name** | **Proposed System** | **Advantages** | **Disadvantages** |
| 1 | Mande Shen, Hansheng Lei, “Improving OCR Performance with Background  Image Elimination” | They present a novel and cost-effective image pre-processing method to accomplish the task. They first enhance the document images before OCR by utilizing the brightness and chromaticity as contrast parameters. Then they convert color images to gray and threshold it. This way, background images can be removed effectively without losing the quality of text characters. | Effective in removing the  background image and thus enhances the performance of  OCR. | Problem of black-white backgrounds. |
| 2 | Vivek Shrivastava and Navdeep Sharma, “ARTIFICIAL NEURAL NETWORK BASED OPTICAL  CHARACTER RECOGNITION” | For the recognition to be accurate, certain topological and geometrical properties are calculated, based on which a character is classified and recognized. Also, the Human psychology perceives characters by its overall shape and features such as strokes, curves, protrusions, enclosures etc. These properties, also called Features are extracted from the image by means of spatial pixel-based calculation. | Less Time Complexity, Very Small Database and High  Adaptability to untrained inputs, with only a small number of features to calculate. | Unable it to recognize stylized  fonts also |
| 3 | Rokus Arnold, Poth Miklos, “Character Recognition Using Neural Networks’ | Neural networks are commonly used to solve sample-recognition problems. One of these is character recognition. The solution of this problem is one of the easier implementations of neural networks. With the help of Matlab’s Neural Network Toolbox, they tried to recognize printed and handwritten characters by projecting them on different sized grids (5x7, 7x11, and 9x13). | Less Time Complexity, Good Extraction Result | Problem of black-white backgrounds. |
| 4 | Vedgupt Saraf, D.S. Rao, “Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency” | The idea of genetic algorithm comes from the fact that it can be used as an outstanding means of combining various styles of writing a character and generates new styles. Closely observing the ability of human mind in the recognition of handwriting, they find that humans are able to recognize characters even though they might be seeing that style for the first time. | Proposed system obtained 98.78% recognition accuracy. | Unable it to recognize stylized  fonts also |
| 5 | P.Murugeswari, Dr.D.Manimegalai, “Complex Background and Foreground Extraction in Color  Document Images using Interval Type-2 Fuzzy” | In this work a new interval type-2 fuzzy based thresholding method is proposed for processing color document images. The proposed method is experimented with varying background of multiple colors and texture and foreground text in any color, font, size and orientation. | The proposed method has automatic parameter estimation so there is no need to tune manually and it can deal with any background complex documents. | Not work on degraded document binarization using the Type-2 fuzzy. |

**Motivation**

This module can be used for text recognition in and give output data which are in computer understandable form. So this project is as important to individual as much to public too.

**Objectives**

* To use document image for processing
* To recognize the characters from document image
* To extract the all characters from document image
* To use OCR process for document processing

**Proposed System**

In proposed system we work on document image to extract the characters from input document image. We first use the document images and preprocess this. That is we convert color images to gray and threshold it. This way, background images can be removed effectively without losing the quality of text characters. Using OCR will successfully extract all the characters from image. This will recognized all the characters of document image and display that in text area which is computer understandable form. So this project is as important to individual as much to public too.

**Proposed System Architecture**

****

**Fig: Proposed System Architecture**

**Algorithm:**

Optical character recognition is usually abbreviated as OCR. It includes the mechanical and electrical conversion of scanned images of handwritten, typewritten text into machine text. It is common method of digitizing printed texts so that they can be electronically searched, stored more compactly, displayed on line, and used in machine processes such as machine translation, text to speech and text mining.

In recent years, OCR (Optical Character Recognition) technology has been applied throughout the entire spectrum of industries, revolutionizing the document management process. OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of OCR, people no longer need to manually retype important documents when entering them into electronic databases. Instead, OCR extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time.

**Steps:**

1. Preprocessing
2. Document and Layout Analysis
3. Paragraphs
4. Lines
5. Recognition
6. Verification

**Hardware/ Software Requirements**

**Hardware Specification:**

1. System: Pentium IV 2.4 GHz.
2. Hard Disk: 40 GB
3. Floppy Drive: 44 Mb.
4. Monitor: 15 VGA Color

**Software Specification:**

1. Operating system: Window 7,8,10.
2. Coding Language: JDK 7.
3. Database: MYSQL 5
4. IDE: Eclipse Luna.

**Conclusion**

In this work we will propose and discussed method of text recognition from document image. The OCR is a wide area for researcher in pattern recognition. A lot of research work has been done and is still being done in OCR for various languages. Each stage of optical character recognition has its own significance and should be designed properly for better results. Will successfully extract the all characters from document image and display it.

**References**

[1] Mande Shen, Hansheng Lei, “Improving OCR Performance with Background Image Elimination”, 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

[2] Vivek Shrivastava and Navdeep Sharma, “ARTIFICIAL NEURAL NETWORK BASED OPTICAL CHARACTER RECOGNITION”, Vol.3, No.5, October 2012.

[3] Rokus Arnold, Poth Miklos, “Character Recognition Using Neural Networks’, 11th IEEE International Symposium on Computational Intelligence and Informatics • 18–20 November, 2010

[4] Vedgupt Saraf, D.S. Rao, “Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency”. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, April 2013.

[5] P.Murugeswari, Dr.D.Manimegalai, “Complex Background and Foreground Extraction in Color Document Images using Interval Type-2 Fuzzy”. International Journal of Computer Applications (0975 – 8887) Volume 25– No.2, July 2011