**TEXT IMAGE SPELL CHECKER**

2010103593 S.S. Sanjay

[sanjaysan01@gmail.com](mailto:sanjaysan01@gmail.com)

2010103563 S. Balaji

[balajisrinivas12@gmail.com](mailto:balajisrinivas12@gmail.com)

**ABSTRACT:**

The main aim of the system is to check for the spelling of a word in an image. To accomplish this, the system extracts the text from the image and then check for its spelling. If the word is spelled correctly, then the system doesn’t do anything. In case of wrong spelling, the system corrects the spelling and creates a new image which has the corrected spelling.

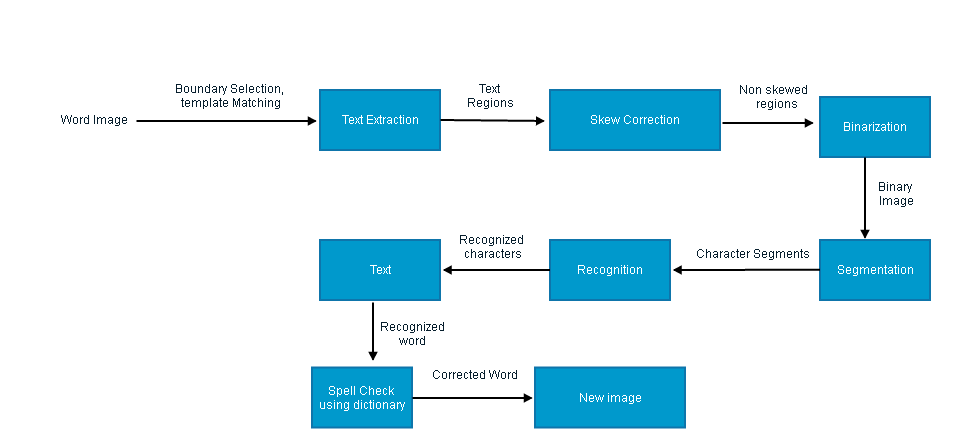
1. **INTRODUCTION:**

Optical Character Recognition, usually abbreviated to OCR, is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text. It is a common method of digitizing printed texts so that they can be electronically searched, stored more compactly, displayed online, and used in machine learning processes such as machine translation, text-to-speech and text mining. OCR is a field of research in Pattern Recognition, AI and Computer Vision.

1. **RELATED WORK:**

* Design of an Optical Character Recognition System for Camera based Handheld Device -Presents an OCR system for camera captured image/graphics embedded textual documents for handheld devices [1].
* An Introduction to the Process of Optical Character Recognition- Presents overview and techniques used for feature extraction that helps in efficient classification of alphabets and numbers of the English Language [2].
* Machine-Printed and hand-written text lines identification- Identify Machine printed and handwritten text of Bangla and Devanagari scripts and classify them as either Machine printed or Hand written [3].

**C. METHODOLOGY:**

**BLOCK DIAGRAM:**

**( I ) Main Idea:**

The main idea behind the system is to recognize the characters from the image and correct the spelling of the extracted word. To explain briefly, the text is first extracted from the image and is then binarized into a black and white image. The edges are then detected from the image using Canny’s Edge Detection Algorithm. The characters are then segmented by

tracing their boundaries from the edge detected image and then using predefined

patterns, the characters are recognized. The word is then formed and is checked with a dictionary to find the closest matching word. If found, the corrected word is displayed and a new image is created with the corrected word.

**(II) Description:**

•**Text Region Extraction Module**: The colour image is converted into a grayscale image by converting the (R,G,B) values in each pixel into a grayscale pixel by applying the transformation: f(x,y) = 0.299 x 𝑟(𝑥,𝑦) +0.587 x 𝑔(𝑥,𝑦)+0.114 x b(x,y).

•**Skew Correction Module**: Camera captured images often suffer from skew and perspective distortion due to unparallel axes and/or planes at the time of capturing the image. So, these text regions are de-skewed using a computationally efficient and fast skew correction technique.

• **Binarization Module**: Binarization is the process of converting gray image into black-white image. A skew corrected text region is binarized using a improved version of Bernsen’s binarization method.

• **Text Segmentation Module**: After binarizing a text region, the horizontal histogram profile {𝑓𝑖, = 1,2, . . ,𝐻𝑅} of the region is analyzed for segmenting the region into text lines. Using vertical histogram profile of each individual text lines, words and characters are segmented.

• **Character Recognition Module**: This is the main algorithm of OCR; an image of every character must be converted to appropriate character code. The sequential steps required to classify an individual binarized character are

* Boundary Selection
* Normalization
* Template Matching
* Character Recognition

• **Spell Checker Module**: In this module , the recognized word is taken and a connection is established with spellcheck.net server and the recognized word is sent to it as part of query. If the spelling is right, the user will not get any error message. But if the spelling is wrong, then the wrong spelling is corrected by comparing the wrong word with the dictionary and then using Levensthein’s Edit Distance to find out the closest matching word. A new image is then created with the corrected word.

**C . EVALUATION:**

**Dataset & Input:**

The dataset for the system is a JPEG image which has one word in Calibri font in black color. A sample of the input and the corresponding output are given below:

**Case 1:**

**Input:**



**Output:**



**Case 2:**

**Input:**



**Output:**



**Intermediate output:**

The intermediate output consists of binarized image of the corresponding input image. It is obtained using Canny’s edge-detection algorithm.

**E. RESULTS AND DISCUSSION:**

The evaluation of the project is done using the correctness of the word extracted from the image by considering various possible words of the font Calibri ,words with font of Calibri family like Calibri bold, Calibri italic and Calibri bold italic.

|  |  |
| --- | --- |
| INPUT | EVALUATION |
| 1)Normal Calibri | The word is correctly recognized as briidge and the word closest to the misspelt word is given as bridge.The actual outcome is the expected outcome |
| 2)Bold Calibri | The word is correctly recognized as briidge and the word closest to the misspelt word is given as bridge.The actual outcome is the expected outcome |
| 3)Italic Calibri | The word is not recognized correctly. There are some letters misconceived by the software. For example,o is recognized as 0.Since the word is not recognized correctly, correct spelling is not predicted |
| 4)Bold Italic Calibri | The word is not recognized correctly. There are some letters misconceived by the software. For example,o is recognized as 0.Since the word is not recognized correctly, correct spelling is not predicted |

**F. LIMITATION AND FUTURE DIRECTION:**

The limitation of the system is that it works only for an image containing a single word of font type “Calibri”. It does not recognize other font types. It does not work for documents containing paragraphs of text. The detection of edges does not work 100% if the text is italicized. But there is a lot of scope for improvement. The future work on this system is to make it recognize fonts related to the “Calibri” family and also to make Edge Detection part work for text which have both bold and italic characteristics.

**G. CONCLUSION:**

The software that is developed aims at recognizing a word’s spelling and check for the existence of the word by looking up a local dictionary. The actual outcome meets the expected outcome in the normal cases where the text is non-italicized text. When the text is italicized, some better edge-detection algorithms must be used to recognize the individual characters of the word. There is still scope for improvement. The text-image spell checker can be used in a variety of scenarios like captcha systems.

**H. REFERENCES:**

[1] Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu and Mita Nasipuri,” Design of an Optical Character Recognition System for Camera based Handheld Device”, 2011.

[2] Umal Patel,”An Introduction to the Process of Optical Character Recognition”, 2013.

[3] U.Pasl, B.B. Chaudhuri,”Machine-printed and hand-written text lines Identification”, 2011.