TEXT EXTRACTION FROM IMAGE APPLICATIONS

K.SUPARNA[1], VADDADI KALKI BHARATHI [2]

[1] Assistant Professor MCA, DEPT, Dantuluri Narayana Raju College , Bhimavaram, Andhrapradesh
Email id:- suparnakalidindi@gmail.com
[2]PG Student of MCA, Dantuluri Narayana Raju College , Bhimavaram, Andhrapradesh
Email id :- vaddadikalkibharathi9@gmail.com

ABSTRACT

Images and movies are increasing on webs and in databases. It is a pressing venture to improve effective techniques to control and retrieve these multimedia resources via their content. The vital object that is included for this task is the text which includes high-level semantic information. When a machine generated textual content is printed against easy backgrounds, it can be transformed to a pc readable structure (ASCII) the usage of modern optical personality consciousness (OCR) technology. However, textual content is frequently printed against shaded or textured backgrounds or is embedded in images. Examples include maps, photographs, advertisements, videos, etc. Current record segmentation and consciousness technologies can't handle these conditions well. Our device takes gain of the distinctive characteristics of text that make it stand out from other photo fabric i.e. textual content possesses sure frequency and orientation information; textual content indicates spatial cohesion—characters of the identical textual content string (a word, or phrases in the identical line) are of comparable, heights, orientation,and,spacing.

1. INTRODUCTION

Today the most data is reachable both on paper or in the shape of snap shots or videos. Large information is stored in images. The cutting-edge science is limited to extracting text towards smooth backgrounds. Thus, there is a need for a system to extract textual content from common backgrounds.

[1] There are quite a number applications in which textual content extraction is useful. These applications encompass digital libraries, multimedia systems, Information retrieval systems, and Geographical Information systems.

[2] The function of textual content detection is to locate the photo areas containing solely text that can be without delay highlighted to the person or fed into an optical persona reader module for recognition.In this paper a new system is proposed which extracts textual content in images. The device takes coloured pix as input. It detects text on the groundwork of positive textual content features: textual content possesses positive frequency and orientation information; textual content shows spatial cohesion—characters of the equal textual content string (a word, or Words in equal line) are of similar heights, orientation and spacing. The photo is then cleaned up so that the textual content stands out.

1.1 Motivation for the work

Text recognition and extraction is needed when the information should be readable both to humans and to a machine and alternative inputs cannot be predefined. The basic Text extraction system was invented to convert the data available on papers in to computer process able documents, so that the documents can be editable and reusable. Traditional techniques are typically multi-stage processes. For example, first the image may be divided into smaller regions that contain the individual characters, second the individual characters are recognized, and finally the result is pieced back together. A difficulty with this approach is to obtain a good division of the original image.

Though tremendous strides have been made in character recognition but it is still considered to be a difficult problem when the data is rotated and non-uniform in scale. We have seen that very few works have been done for Indian languages using Fuzzy logic. In this work, we have taken the problem of improving the recognizing capability of compound characters using Fuzzy logic so as to achieve accurate character values.

1.2 Problem Statement

As we can see in our daily lives, people take images of some documents when they have no other source to take that document with them, but later they have to read each and every word from it. So, we thought to make a project in which we can just take an image and process it to extract the text present in the image. It saves a lot of time to read the text from an image.

2. LITERATURE SURVEY AND RELATED WORK

This material serves as a guide and update for readers working in the Character Recognition area. Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu, and Mita Nasipuri(2011) [1] presents a complete Optical Character Recognition(OCR) system for camera captured image textual documents for handheld devices. Firstly, text regions are extracted and skew corrected. Then, these text regions are binarized and segmented into lines and characters. Characters are passed into the recognition module. Pranob K Charles, V.Harish, M.Swathi(2012)[2] describes the techniques for converting textual content from a paper document into machine readable form. The computer actually recognizes the characters in the document through a revolutionizing technique called Optical Character Recognition. Chirag Patel ,Atul Patel, Dharmendra Patel(2012) [3] recognize the characters in a given scanned documents and study the changes in the Models of Artificial Neural Network. It describes the behaviors of different Models of Neural Network used in Optical Character Recognition.

Image processing is analysis and manipulation of a digitized image, so as to enhance its quality with the help of mathematical operations by using any kind of signal processing where the input is a picture or an image or a video frame. The output of image processing will be either a picture or set of characters or parameters associated with the given input image. This is a set of computational techniques for analyzing, enhancing, compressing and reconstructing image.

Image Processing is set of computational techniques for analyzing, enhancing, compressing, and reconstructing images. Its main components are importing, in which an image is captured through scanning or digital photography; analysis and manipulation of the image, accomplished using various specialized software applications; and output. Image processing has extensive applications in many areas, including astronomy, medicine, industrial robotics, and remote sensing by satellites.

Image Processing provides a comprehensive set of reference-standard algorithms and workflow apps for image analysis, visualization, and algorithm development. Image Processing can interactively segment image data, compare image registration techniques, and batch-process large data sets.

There are various kinds of techniques for processing an image like linear scaling, optical methods, fuzzy techniques, digital processing.

Image processing generally involves three steps:

- Importing and Loading the image by using image acquisition tools.
- Analyzing and manipulating image to extract the information.
- Output the result. The result might be the image or a picture altered in some way or it may be a report based on analysis of the image.

3. EXISTING SYSTEM

Text extraction from image applications involves the development of software and systems that can automatically extract text content from images, making it accessible and searchable. This technology has various applications in fields such as document management, OCR (Optical Character Recognition), data extraction, and more.The existing system for text extraction from image applications typically consists of several key components and functionalities.

Image Input:

Users can upload images in various formats, including JPEG, PNG, TIFF, and PDF.

Image Preprocessing:

Image preprocessing techniques are applied to enhance the quality of the input image.

This may include noise reduction, contrast adjustment, and resizing to optimize OCR accuracy. Optical Character **Recognition (OCR):**

OCR engines are employed to analyze the processed image and convert any text content into machine- readable text. Leading OCR engines include Tesseract, ABBYY FineReader, and Google Cloud Vision API.

Language Detection:

Language detection algorithms may be integrated to identify the language of the extracted text, which can aid in further and analysis.

4. PROPOSED SYSTEM

Optical Character Recognition (OCR), is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. It is the mechanical or electronic conversion of images of typewritten or printed text into machine encoded text. Images captured by a digital camera differ from scanned documents or image-only PDFs. They often have defects such as distortion at the edges and dimmed light, making it difficult for most OCR applications, to correctly recognize the text. The latest version of ABBYY Fine Reader supports adaptive recognition technology specifically designed for processing camera images. It offers a range of features to improve the quality of such images, providing you with the ability to fully use the capabilities of your digital devices. A common problem faced by travelers is that of understanding unfamiliar language. Failing to understand unknown languages, when travelling can lead to minor problems. These systems are usually composed of two subsytems that perform text extraction and text translation respectively. The extraction and translation parts are relatively well developed and there exist a large variety of software packages or web services that perform these tasks. The challenge iswithextractingtheexacttextfromthe. Images and translating it to known language. In a typical scenario, a user takes a picture of a text area with the cell phone camera, the text is extracted from the image. In image processing and computer vision, edge detection treats the localization of significant variations of a gray level image and the identification of the physical and geometrical properties of objects of the scene. The variations in the gray level image commonly include discontinuities (step edges), local extreme (line edges) and junctions. Most recent edge detectors are autonomous and multiscale then include three main processing steps smoothing, differentiation and labeling. The edge detectors vary according to these processing steps, to their goals, and to their mathematical and computational complexity. The extracted text is then translated using translation engine which contains the database of languages. Then the translated text is given as output. The Purpose of this project is to implement text extraction from the image and translating the given text. Many different methods are used for extracting the text from the images. Properties like color, intensity, edges etc are related in extracting the text To carry out this task we have four modules those are Image Capture, Text Identification, Language conversion, PDF generation. A mobile camera is used to capture the image.

5. METHODOLOGIES

Pre-processing Module

The paper document is generally scanned by the optical scanner and isconvertedin to the form of a picture. At this stage we have the data in the form of image and this image can be further analysed so that's the important information can be retrieved. The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique forthresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a pre-processor to smooth the digitized characters. The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. The most common techniques for smoothing, moves a window across the binary image of the character, applying certain rules to the contentsof the window. So, to improve quality of the input image, few operations are performedfor enhancement of image such as noise removal, normalization, binarization etc.

Noise Removal

Image noise is an unavoidable side-effect occurring as a result of image capture, more simply understood as inaudible, yet inevitable fluctuations. In a digital camera, if the light which enters the lens misaligns with the sensors, it will create image noise. Even if noise is not so obviously visible in a picture, some kind of image noise is bound to exist. Every type of electronic device receives and transmits some noise and sends it onto what it is creating. When the images are transmitted over channels, they are corrupted with impulse noise due to noisy channels. This impulse noise consists of large positive and negative spikes. The positive spikes have values much larger than the background and thus they appear as bright spots, while the negative spikes have values smaller thanthe background and they appear as darker spots. Both the spots for the positive and negative spikes are visible to the human eye. Also, Gaussian type of noise affects the image.

Filtering

Filters are required for removing noises before processing. There are lots of filtersin the paper to remove noise. They are of many kinds as linear smoothing filter, medianfilter, wiener filter and Fuzzy filter. In this filtering technique, the three primaries (R, Gand B) are done separately. It is followed by some gain to compensate for attenuation resulting from the filter. The filtered primaries are then combined to form the colouredimage. This process is very simple. This approach shown in figure below as.

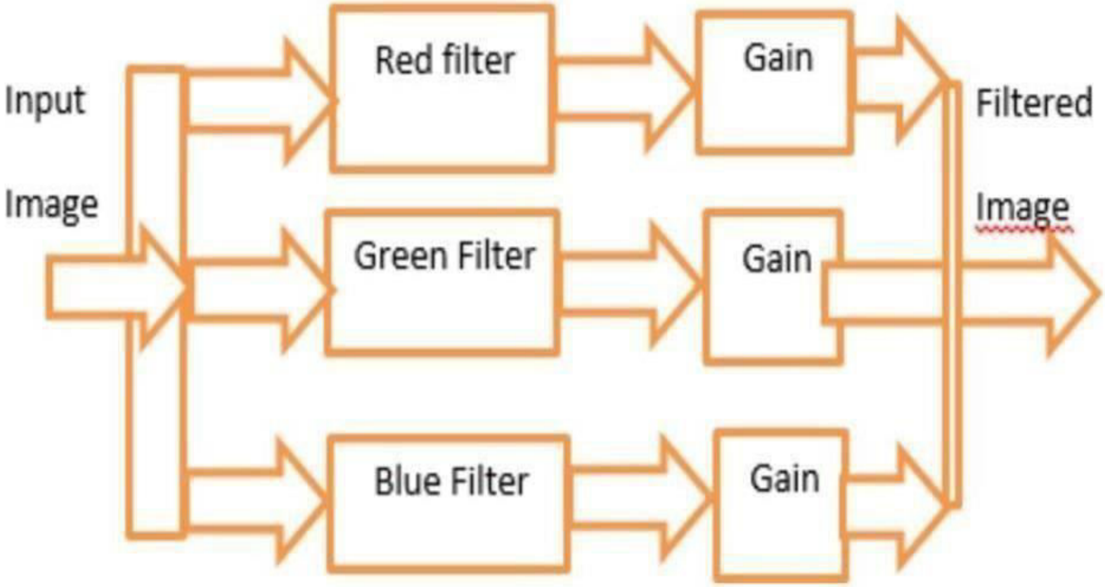
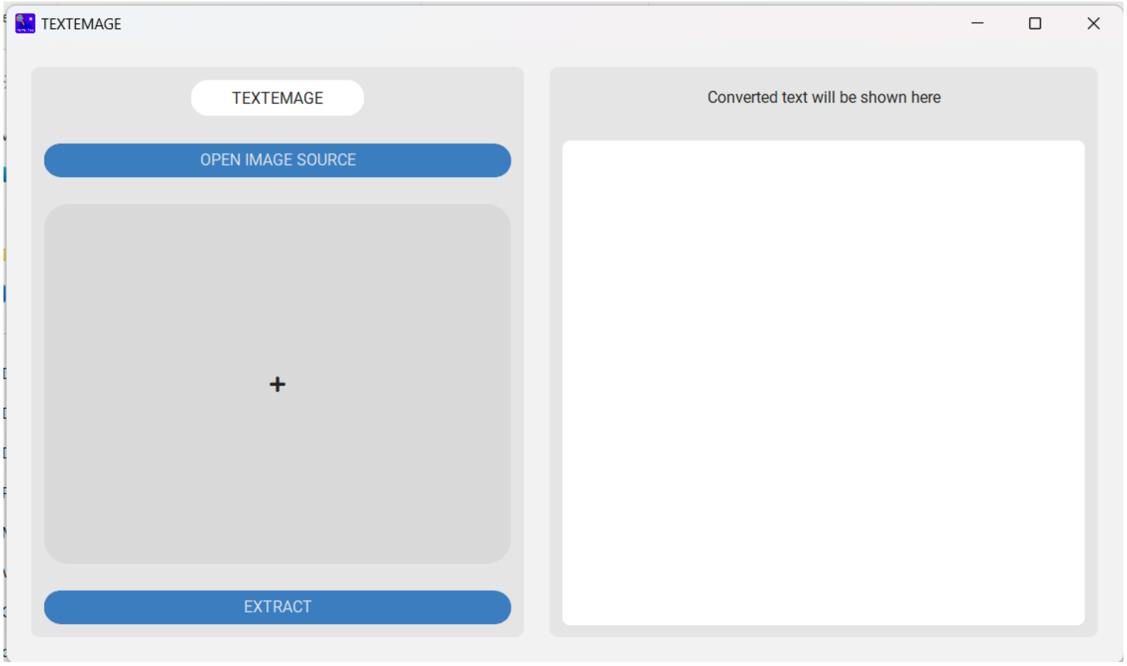


Fig.1 - Filtering

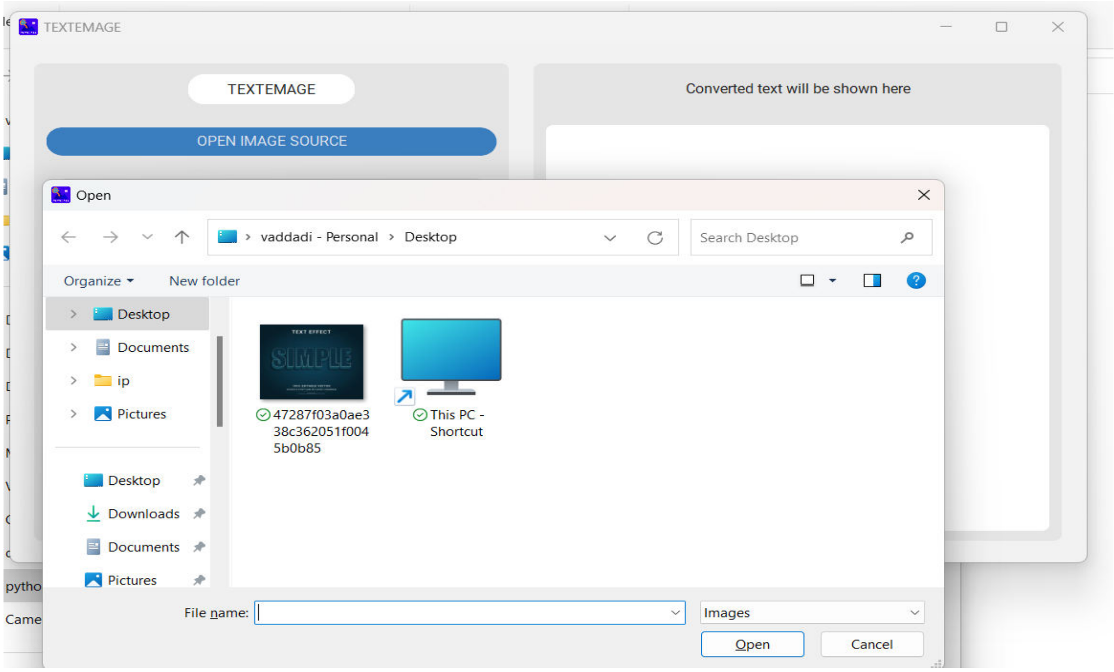
6. RESULTS AND DISCUSSION SCREEN SHOTS

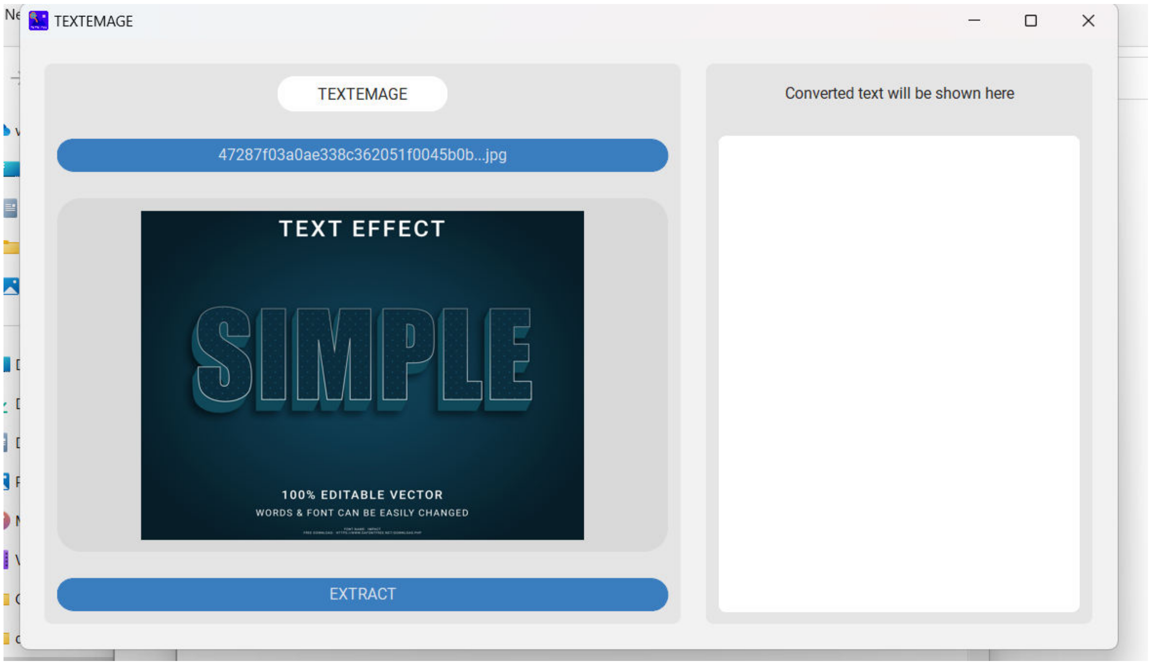
OUTPUT

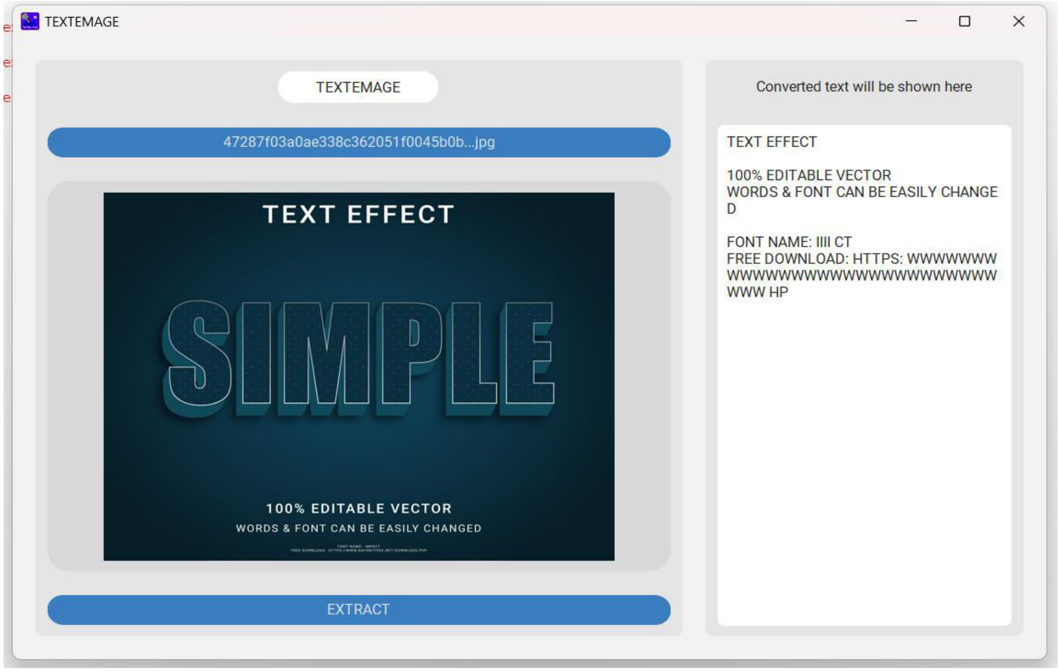
SCREENSHOTS



Converted text will be shown here


TEXTEMAGE
TEXTEMAGE
OPEN IMAGE SOURCE
Converted text will be shown here
Open
vaddadi - Personal > Desktop
Search Desktop
Organize
New folder
Desktop
Documents
ip
Pictures
Desktop
Downloads
Documents
Pictures
47287f03a0ae338c362051f0045b0b85
This PC - Shortcut
File name:
Images
Open
Cancel


TEXTEMAGE
TEXTEMAGE
47287f03a0ae338c362051f0045b0b...jpg
TEXT EFFECT
SIMPLE
100% EDITABLE VECTOR
WORDS & FONT CAN BE EASILY CHANGED
EXTRACT
Converted text will be shown here


TEXTEMAGE
TEXTEMAGE
47287f03a0ae338c362051f0045b0b...jpg
TEXT EFFECT
SIMPLE
100% EDITABLE VECTOR
WORDS & FONT CAN BE EASILY CHANGED
EXTRACT
Converted text will be shown here

```
TEXT EFFECT

100% EDITABLE VECTOR
WORDS & FONT CAN BE EASILY CHANGE
D

FONT NAME: IIII CT
FREE DOWNLOAD: HTTPS: WWWWWWW
WWWWWWWWWWWWWWWWWWWWWWWW
WWW HP
```

7. CONCLUSION AND FUTURE SCOPE

CONCLUSION:

This is the discussion about optical character recognition techniques to translate the text from unknown language text into known language.The system has the capability to recognize characters with accuracy exceeding 90% mark. The advantage of this system is that it is easily portable and its scalability which can recognize various languages and also help in translating the text in different languages. The accurate recognition is directly depending on the nature of the material to be read and by its quality.

FUTURE SCOPE :

As the proposed system can only organize text for some of the image documents such as identification proofs and a specific application form, this can be extended to other image documents as per user request and test it on well-designed datasets for improving the accuracy of the system.

8. REFERENCES

1. . [1] Victor Wu, R. Manmatha, Edward M. Riseman.:"Text
2. Finder: An Automatic System to Detect and Recognise
3. Text in Images."
4. [2] Miriam León, AntoniGasull.:"Text Detection in Images and
5. Video Sequences."
6. [3] Sneha Sharma.:"Extraction of Text Regions in Natural
7. Images."
8. [4] KhaledHammouda, Prof. ED Jernigan.:"Texture
9. Segmentation using Gabor Filters."
10. [5] HaidarAlmohri, John S. Gray, HishamAlnajjar.:" A real-time
11. dsp-based Optical Character Recognition System for
12. isolated arabic characters."
13. [6] AmerDawoud and Mohamed Kame1.: "Binarization of
14. Document Images Using Image