# ABSTRACT

Text recognition in images is a research area which attempts to develop a computer system with the ability to automatically read the text from images. These days there is a huge demand in storing the information available in paper documents format in to a computer storage disk and then later reusing this information by searching process. One simple way to store information from these paper documents in to computer system is to first scan the documents and then store them as images. But to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word. So, with the help of Fuzzy Logic we extract Editable Text from an Input Image.

**Keywords:** Text Recognition, Text Extraction, Fuzzy Logic, Pre-processing, Segmentation, Feature Extraction**.**

# CONTENTS

# LIST OF FIGURES

# 1. INTRODUCTION

## 1.1 Introduction

Today the most information is available either on paper or in the form of photographs or videos. Large information is stored in images. The current technology is restricted to extracting text against clean backgrounds. Thus, there is a need for a system to extract text from general backgrounds. There are various applications in which text extraction is useful. These applications include digital libraries, multimedia systems, Information retrieval systems, and Geographical Information systems. The role of text detection is to find the image regions containing only text that can be directly highlighted to the user or fed into an optical character reader module for recognition.

The information from these image documents would give higher efficiency and ease of access if it is converted to text form. The process by which Image Text converted into plain text that computer can recognize its ASCII character is Text Extraction. The information from image documents should be converted into text in order to get efficient use and access of it like archiving or reporting that are used in different image-based applications such as office works.

Document papers that need to be digitized and used for archiving or indexing or information retrieval process are increasingly common today, for example scanned documents of office works, in magazines, advertisements and web pages. Robust and efficient extraction of text from these documents is a challenging problem due to different properties of text in image.

Textual data present in the images contain useful information for indexing and automatic annotations. Extraction of this useful information involves text detection, localization of text, classification, and then recognition of text. Fuzzy logic determines the degree of truth values. This logic helps to identify and match the characters accurately with trained data.

## 1.2 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.

## 1.3 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.

## 1.4 Organization of the thesis

Our thesis is organized as follows.

Chapter 2: This chapter provides the literature survey of the project and we provide the necessary background.

Chapter 3: We present our proposed architecture, method and the assumptions on which we based our approach.

Chapter 4: Describes the design of the project.

Chapter 5: Deals with the experimental setup that we used and we provide the results of the experiments.

Chapter 6: We conclude and we suggest possible expansions of our research.

## 2.1 Image Processing

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

## 2.3 OCR

OCR (Optical Character Recognition) is the use of technology to distinguish printed or handwritten text characters inside digital images of physical documents, such as a scanned paper document. The basic process of OCR involves examining the text of a document and translating the characters into code that can be used for data processing. OCR is sometimes also referred to as text recognition.

OCR systems are made up of a combination of hardware and software that is used to convert physical documents into machine-readable text. Hardware, such as an optical scanner or specialized circuit board is used to copy or read text while software typically handles the advanced processing. Software can also take advantage of artificial intelligence (AI) to implement more advanced methods of Intelligent Character Recognition (ICR), like identifying languages or styles of handwriting.

The major OCR technology providers began to tweak OCR systems to deal more efficiently with specific types of input. Beyond an application-specific lexicon, better performance may be had by taking into account business rules, standard expression, rich information contained in colour images. This strategy is called "Application-Oriented OCR" or "Customized OCR", and has been applied to OCR of invoices, screenshots, ID cards, driver licenses, and automobile manufacturing.

## 2.4 Tesseract

Tesseract is an optical character recognition engine for various operating systems. It is used for processing of image to recognize text. In tesseract the recognized characters are stored as a variable to compare with trained data. Tesseract is an open source optical character recognition engine. It was developed at HP in between 1984 to1994. It was modified and improved in 1995 with greater accuracy. In late 2005, HP released Tesseract for open source. It is highly portable. It is more focused towards providing less rejection than accuracy. Currently only command base version is available. It provides support for various languages.

Tesseract OCR works in step by step manner i.e. First step is Adaptive Thresholding, which converts the image into binary images. Next step is connected component