Informatics Institute of Technology

In Collaboration with

University of Westminster (UOW)

A System to Torn Document Analysis and Reconstruction

A dissertation by

W1582956 | Kaveesha Chethiyawardena | 2015052

Supervised by

Mr Rathesan Sivagnanalingam

Submitted in partial fulfillment of the requirements for the

Bsc (Hons) Software engineering degree

Department of Computing

May 2019

Abstract

The process of preserving information and data is one of the most important tasks carried out by human beings since early evolutionary period. Since then many different methods were practiced and the oldest and the most prominent way that remained till present days is documenting of valuable information. But with the time, climatic conditions, natural and intentional activities like crimes and frauds, the documents containing valuable information tends to damage and tear which leads to loss of valuable documents. As natural paper that use documentation is biodegradable the tearing and damaging cannot be avoided but they can be reconstructed as same as the original document which will prevent the loss of valuable information and misleading of the society. The existing solutions of torn document reconstruction is mainly focused on edge analysis, shape detection, texture analysis, corner and segment matching, snippets analysis and pattern analyzing but the content within torn document pieces analysis and identification wasn't paid the required attention. Therefore, the aim of this project is to research and implement a system which could reconstruct torn documents by edge detection and content analysis to predict unclear words and correct grammar errors preventing loss of the important documents. To accomplish this aim documents that are torn in to different shapes with irregular edges and different content was selected. The Canny edge detection, Brute Force approach with NLP and OCR text identification approaches were used to implement the proposed system.

Keywords – Image Processing, Torn Document, NLP, OCR, Python