Testing the extraction

Tan Shu Hui

22004830

shuhui@email.com

Abstract—Digital images are getting popular rapidly. Every day, many images have been generated by many groups like students, engineer, doctors, according to their varying needs. They can access images based on its primitive features or associated text. Text present in such images can provide meaningful information. We aim to retrieve the content and summarize the visual information automatically from images. Optical character recognition system that involves several algorithms are required for this purpose. Tesseract is currently the most accurate optical character recognition engine which was developed by HP Labs and is currently owned by Google. In this paper, we extract text from images using text localization, segmentation and binarization techniques. Text extraction can be achieved by applying text detection that identifies image parts containing text, text localization finds the exact position of the text, text segmentation separates the text from its background and binarization process converts the coloured images into binary. On this binary image, character recognition is applied to convert it into ASCII text. Text extraction is used in creating e-books from scanned books, image searching from a collection of visual data etc.

INTRODUCTION

Due to rapid development in digital technology, we have a large collection of information stored in the form of images [8] resulting in digitization of resources in various industries. Recent studies have been conducted on image processing that shows the importance of content retrieval from images. Text extraction from images and converting them to ASCII text can be achieved using optical character recognition (OCR) systems. OCR is helpful and vital in various applications including digital libraries, information retrieval systems, multimedia systems and Geographical Information systems [1]. OCR system almost reduced the keyboard interface between man and machine and helps in automation of office which saves a lot of time and human effort [3,27]. The accuracy of an OCR system is sometimes dependent on text pre-processing and segmentation methods. Level of difficulty in extraction of text

is dependent on different styles, size, complex image background, orientation etc [4]. Various methodologies have been used to perform text extraction from images such as text detection, text localization, text segmentation etc. Text detection plays a very important role in determining and highlighting image parts containing only text [1,28]. The images captured in OCR systems often include skew and perspective distortions due to some human errors which are also required to be removed. Binarization of skew corrected images using simple yet efficient binarization techniques that are performed before the segmentation process [14]. After processing of input image, we obtain a binarized image i.e. texts are written in black colour on white background. On this binarized image, text localization is performed that involves separating characters from each word in the image by scanning pixels sequentially [6]. Sometimes situation occur such that components of adjacent characters are touched or overlapped which creates difficulties in the segmentation task. This problem occurs frequently due to modification of upper and lower zones; thus, it is an important stage [13].
Tesseract is an open source software that helps in text extraction with comparatively high accuracy when compared with other OCR systems. Tesseract does not have its own page layout, therefore it assumes to have input as binary image and optionally defined polygonal text regions [2]. Processing period followed is a step-by-step pipeline of connected component analysis (called Blobs [2]) that recognizes the text as black-on white text, organization of blobs into text lines, breaking text lines into words according to the character spacing kinds and so on [16]

II. RESEARCH HISTORY

For past few years, the need for retrieving data from images and storing them for future reference has rapidly increased. Several researches have been performed to study the various approaches that could be helpful in extracting data from images. These approaches include methods [1] related to various extraction processes such as text detection, text localization, text segmentation [8] etc. Additionally, various properties of an image such as colour, intensity, component connection, orientation, text-style [8] etc are useful in separating the text regions from their backgrounds and other regions of the image [1]. Machine recognition of handwritten texts has been in research for pattern recognition. Previously Tesseract has been used to perform user specific training on both isolated and free-flow text by specifically using lower case

Roman script [3]. Being suspended for more than 10 years,

Tesseract now give base to major commercial engines with improved accuracy. Tesseract was developed by HP, but never used by it. Later it was modified, improved and maintained by Google [4]. Although Tesseract is helpful in extracting data from images with more accuracy, but along with, it brought some flaws with itself like over-segmentation for some characters, and under-segmentation or rejection in case of cursive word segments are few of them [3]. Its unusual choice of features is probably the key strength and instead of raw outlines, using polygonal approximation is its key weakness [2]. Various projects have been developed using Tesseract to implement real-world scenarios related to manuscripts, data extraction and archiving from images, effective manipulation of image databases, language processing and many more.