International Journal of Engineering & Technology, 7 (2.24) (2018) 361-364

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Improving Optical Character Recognition Techniques

Nitin Ramesh, Aksha Srivastava*, K. Deeba

Department of Computer Science and Engineering, SRM University, Chennai, India
**Corresponding Author E-mail: nitin_ramesh@srmuniv.edu.in*

Abstract

Document text recognition uses a concept called OCR (optical character recognition),which is the recognition of printed or written text characters by a computer. This involves scanning a document containing text, and converting character by character to their digital form. Thus, it is defined as the process of digitizing a document image into its constituent characters. Equipment used to obtain clearer images for analysis are cameras and flatbed scanners. Even though it’s been out in the world since 1870, the OCR technology is yet to reach perfection. This demanding nature of Optical Character Recognition has made various researchers, industries and technology enthusiasts to divulge their attention to this field. In recent times one can notice a significant increase in the number of research organizations investing their time and effort in this field. In this research, the progress, different aspects and various issues revolving in this field have been summarized. The aim is to present a scrupulous overview of various proposals, advancements and discussions aimed at resolving various problems that arise in traditional OCR.

Keywords: *Text Recognition, OCR, Image Analysis, Photo Scanning, Scanned Image.*

1. Introduction

Optical Character Recognition (OCR) is a piece of algorithm that involves the conversion of images containing text and printed messages into a digital format that can be handled by a machine.
The human neurological system has been a wonder in the way that it has the capability to make sense out of patterns presented to it in a haphazard manner. It has always taken the lead in this particular field, especially when compared to machines, which trump the brain in mathematical and seemingly normal computations which involve methods that are traditionally binary.
While the brain recognizes photographs and its subjects, along with any kind of written content i.e handwritten or printed, machines aren’t able to appraise the information that is being provided in the image.
As a consequence of which, there have been immense and numerous efforts in this discipline attempting to bridge the gap of incomprehensibility hench transforming an image of a document into a more understandable format when seen by a machine.

2. Types of Character Recognition

Printed

OCR targets typewritten text, one character at a time. This is accomplished using pattern matching and feature analysis.
ICR on the other hand may also target handwritten text and makes use of machine learning techniques.

Handwritten

The offline mode for recognition is the processing of a static document, whereas the online version is much more advanced and uses handwriting movement analysis. Instead of using the commonly used pattern learning algorithm, the online mode allows us to capture motions, i.e the order in which the segments are drawn and their direction of strokes.
Irrespective of the type of character recognition, the documents are converted to a grayscale format. Since this basic grayscale conversion isn’t enough various other processing techniques are used to allow better recognition of characters and letters.
The detection of words and characters dealt by Optical Character Recognition takes place when the data is obtained by an optical medium, which includes devices like camera and scanners.
Optical Character Recognition ideally deals with pixelated data of two types and those types can be divided on the basis of whether the text is handwritten or printed, i.e either handwritten OCR or printed OCR. The symbols and signs can be of any size and orientation as a handheld camera/device won’t provide the accuracy of a flatbed scanner and handwritten text varies from person to person.
Due to the diversity available in types of handwriting, and taking into account various styles across the globe, it is easy to conclude that Handwritten OCR is challenging to implement when compared to Printed OCR in which the images to be processed have a custom style, matching to suit the fonts they have been written in. They represent a more homogenous pattern as compared to the former.

Summarized below are the drawbacks of existing OCR systems:

A. Due to the various constraints that arise like shadows, skewing, blurred lines, this type of conversion requires expensive pre-processing algorithms which can be flawed, before any kind of character extraction or recognition stages can be implemented on the model.
B. The recognition they do is performed on already present data so they are not dynamic or during real time.
C. The accuracy of popular systems such as Tesseract and OCR Space drop drastically when the character height is below 20 pixels.

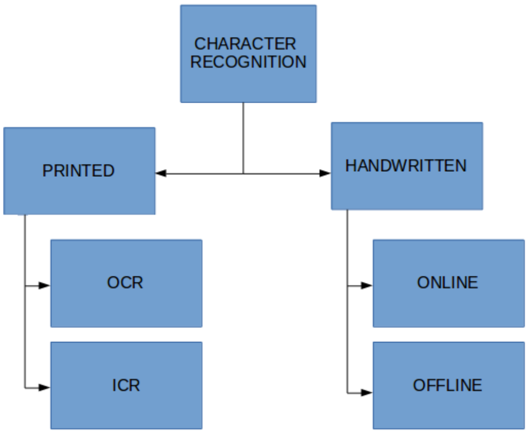


Fig. 2: Various types of OCR

3. Proposed System

The proposed system makes use of Tesseract OCR engine while eliminating various drawbacks, hence improving performance.

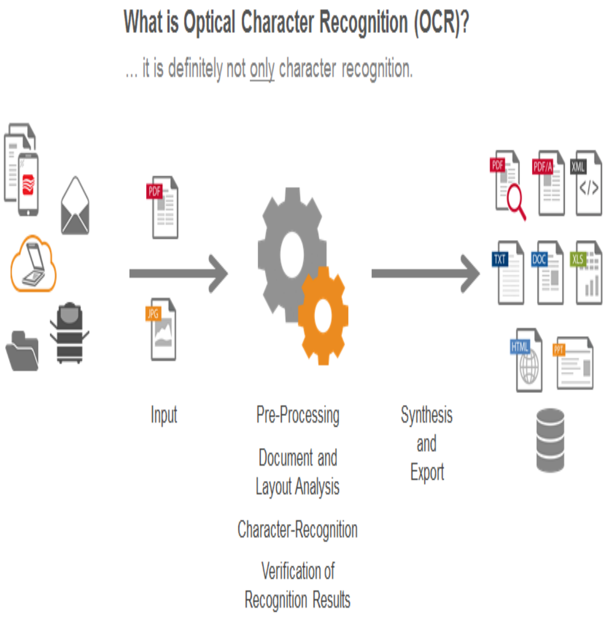
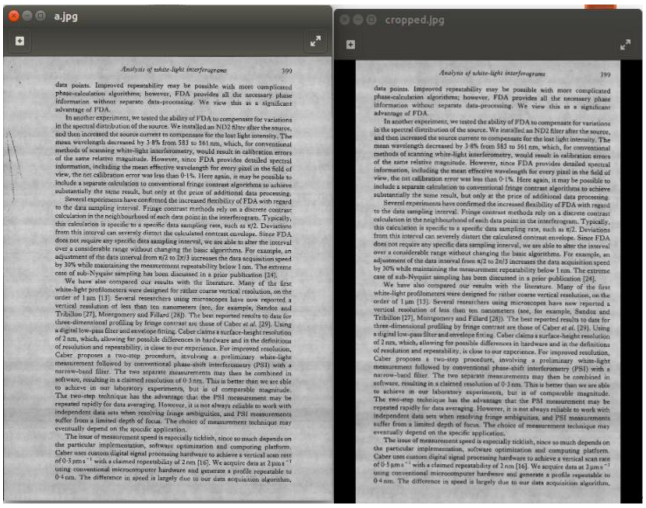

What is Optical Character Recognition (OCR)?
... it is definitely not only character recognition.
PDF
JPG
PDF
PDF/A
XML
TXT
DOC
XLS
HTML
PPT
Input
Pre-Processing
Document and Layout Analysis
Character-Recognition
Verification of Recognition Results
Synthesis and Export

Fig. 3: OCR Procedure

3.1 Pre-Processing

3.1.1 Text Region Extraction



Analysis of white-light interferograms 399

data points. Improved repeatability may be possible with more complicated phase-calculation algorithms; however, FDA provides all the necessary phase information without separate data-processing. We view this as a significant advantage of FDA.

In another experiment, we tested the ability of FDA to compensate for variations in the spectral distribution of the source. We installed an ND2 filter after the source, and then increased the source current to compensate for the lost light intensity. The mean wavelength decreased by 3·8% from 583 to 561 nm, which, for conventional methods of scanning white-light interferometry, would result in calibration errors of the same relative magnitude. However, since FDA provides detailed spectral information, including the mean effective wavelength for every pixel in the field of view, the net calibration error was less than 0·1%. Here again, it may be possible to include a separate calculation to conventional fringe contrast algorithms to achieve substantially the same result, but only at the price of additional data processing.

Several experiments have confirmed the increased flexibility of FDA with regard to the data sampling interval. Fringe contrast methods rely on a discrete contrast calculation in the neighbourhood of each data point in the interferogram. Typically, this calculation is specific to a specific data sampling rate, such as π/2. Deviations from this interval can severely distort the calculated contrast envelope. Since FDA does not require any specific data sampling interval, we are able to alter the interval over a considerable range without changing the basic algorithms. For example, an adjustment of the data interval from π/2 to 2π/3 increases the data acquisition speed by 30% while maintaining the measurement repeatability below 1 nm. The extreme case of sub-Nyquist sampling has been discussed in a prior publication [24].

We have also compared our results with the literature. Many of the first white-light profilometers were designed for rather coarse vertical resolution, on the order of 1 μm [13]. Several researchers using microscopes have now reported a vertical resolution of less than ten nanometers (see, for example, Sandoz and Tribillon [27], Montgomery and Fillard [28]). The best reported results to date for three-dimensional profiling by fringe contrast are those of Caber *et al.* [29]. Using a digital low-pass filter and envelope fitting, Caber claims a surface-height resolution of 2 nm, which, allowing for possible differences in hardware and in the definitions of resolution and repeatability, is close to our experience. For improved resolution, Caber proposes a two-step procedure, involving a preliminary white-light measurement followed by conventional phase-shift interferometry (PSI) with a narrow-band filter. The two separate measurements may then be combined in software, resulting in a claimed resolution of 0·3 nm. This is better than we are able to achieve in our laboratory experiments, but is of comparable magnitude. The two-step technique has the advantage that the PSI measurement may be repeated rapidly for data averaging. However, it is not always reliable to work with independent data sets when resolving fringe ambiguities, and PSI measurements suffer from a limited depth of focus. The choice of measurement technique may eventually depend on the specific application.

The issue of measurement speed is especially ticklish, since so much depends on the particular implementation, software optimization and computing platform. Caber uses custom digital signal processing hardware to achieve a vertical scan rate of 0·5 μm s^{-1} with a claimed repeatability of 2 nm [16]. We acquire data at 2 μm s^{-1} using conventional microcomputer hardware and generate a profile repeatable to 0·4 nm. The difference in speed is largely due to our data acquisition algorithm,

Analysis of white-light interferograms 399

data points. Improved repeatability may be possible with more complicated phase-calculation algorithms; however, FDA provides all the necessary phase information without separate data-processing. We view this as a significant advantage of FDA.

In another experiment, we tested the ability of FDA to compensate for variations in the spectral distribution of the source. We installed an ND2 filter after the source, and then increased the source current to compensate for the lost light intensity. The mean wavelength decreased by 3·8% from 583 to 561 nm, which, for conventional methods of scanning white-light interferometry, would result in calibration errors of the same relative magnitude. However, since FDA provides detailed spectral information, including the mean effective wavelength for every pixel in the field of view, the net calibration error was less than 0·1%. Here again, it may be possible to include a separate calculation to conventional fringe contrast algorithms to achieve substantially the same result, but only at the price of additional data processing.

Several experiments have confirmed the increased flexibility of FDA with regard to the data sampling interval. Fringe contrast methods rely on a discrete contrast calculation in the neighbourhood of each data point in the interferogram. Typically, this calculation is specific to a specific data sampling rate, such as π/2. Deviations from this interval can severely distort the calculated contrast envelope. Since FDA does not require any specific data sampling interval, we are able to alter the interval over a considerable range without changing the basic algorithms. For example, an adjustment of the data interval from π/2 to 2π/3 increases the data acquisition speed by 30% while maintaining the measurement repeatability below 1 nm. The extreme case of sub-Nyquist sampling has been discussed in a prior publication [24].

We have also compared our results with the literature. Many of the first white-light profilometers were designed for rather coarse vertical resolution, on the order of 1 μm [13]. Several researchers using microscopes have now reported a vertical resolution of less than ten nanometers (see, for example, Sandoz and Tribillon [27], Montgomery and Fillard [28]). The best reported results to date for three-dimensional profiling by fringe contrast are those of Caber *et al.* [29]. Using a digital low-pass filter and envelope fitting, Caber claims a surface-height resolution of 2 nm, which, allowing for possible differences in hardware and in the definitions of resolution and repeatability, is close to our experience. For improved resolution, Caber proposes a two-step procedure, involving a preliminary white-light measurement followed by conventional phase-shift interferometry (PSI) with a narrow-band filter. The two separate measurements may then be combined in software, resulting in a claimed resolution of 0·3 nm. This is better than we are able to achieve in our laboratory experiments, but is of comparable magnitude. The two-step technique has the advantage that the PSI measurement may be repeated rapidly for data averaging. However, it is not always reliable to work with independent data sets when resolving fringe ambiguities, and PSI measurements suffer from a limited depth of focus. The choice of measurement technique may eventually depend on the specific application.

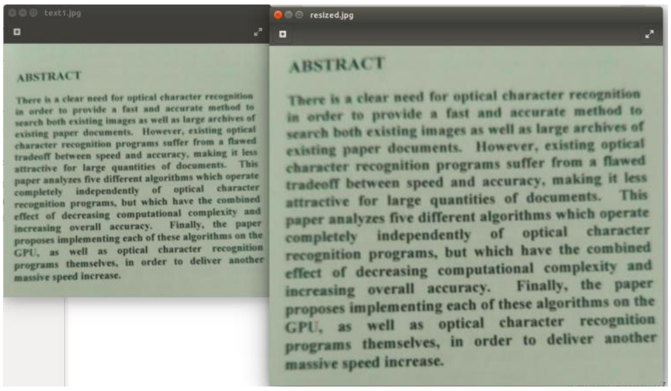
The issue of measurement speed is especially ticklish, since so much depends on the particular implementation, software optimization and computing platform. Caber uses custom digital signal processing hardware to achieve a vertical scan rate of 0·5 μm s^{-1} with a claimed repeatability of 2 nm [16]. We acquire data at 2 μm s^{-1} using conventional microcomputer hardware and generate a profile repeatable to 0·4 nm. The difference in speed is largely due to our data acquisition algorithm,

Fig. 4: Left image is the original. The right image is cropped

In order to extract the text regions, a document populated with words is shrunk to the level where the spaces between words become negligible.

Thus when finding contours, only one contour encompassing all the words is obtained, and the unnecessary area around the paragraph can be cropped.

3.1.2 Image Resizing



ABSTRACT

There is a clear need for optical character recognition in order to provide a fast and accurate method to search both existing images as well as large archives of existing paper documents. However, existing optical character recognition programs suffer from a flawed tradeoff between speed and accuracy, making it less attractive for large quantities of documents. This paper analyzes five different algorithms which operate completely independently of optical character recognition programs, but which have the combined effect of decreasing computational complexity and increasing overall accuracy. Finally, the paper proposes implementing each of these algorithms on the GPU, as well as optical character recognition programs themselves, in order to deliver another massive speed increase.

ABSTRACT

There is a clear need for optical character recognition in order to provide a fast and accurate method to search both existing images as well as large archives of existing paper documents. However, existing optical character recognition programs suffer from a flawed tradeoff between speed and accuracy, making it less attractive for large quantities of documents. This paper analyzes five different algorithms which operate completely independently of optical character recognition programs, but which have the combined effect of decreasing computational complexity and increasing overall accuracy. Finally, the paper proposes implementing each of these algorithms on the GPU, as well as optical character recognition programs themselves, in order to deliver another massive speed increase.

Fig. 5: Left image is the original. The right image is enlarged

Since Tesseract scans the image pixel by pixel, reading an image with an average character height above 20 pixels will increase computation time.

This step also helps eliminating one major flaw of Tesseract, where the OCR accuracy drops for characters below a height of 20 pixels. As shown in Fig.5 the original image has an average character height of 12 pixels, hence the image is resized to the value where the average character height is 20 pixels.

In the example given in Fig.5 the accuracy of a straightforward OCR procedure is increased from 17.54% to 98.24%, by resizing the image accordingly. Thus resolving one of the drawbacks of engines, such as Tesseract.

3.1.3 Binarization and Denoise

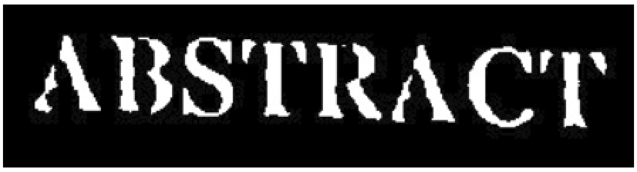
Binarization is the process of conversion of an image to black and white. The high contrast between the background and the characters make OCR more accurate.

The process of binarization is done through thresholding the color image. There are two types of thresholding, one is global thresholding and the other being adaptive thresholding.

Thresholding changes the pixel color value to a minimum if it is below a threshold or to a maximum if it is above the threshold value.

Global thresholding is done by fixing one threshold value for an entire image. This becomes ineffective when the image has different levels of illumination.

Hence we use Adaptive thresholding, which computes a different threshold for small regions of the image. This results in broken characters and noise around the characters, hence binarization is followed by a median blur and dilation, which bridge the spaces.



ABSTRACT

Fig. 6: Binarized image with broken characters

Mask Size = 21 and Subtracting constant = 11

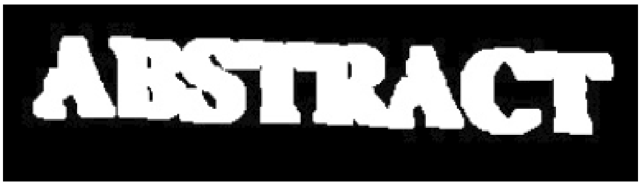


Fig. 7: Median Blur followed by Dilation

A vast amount of noise is removed and the characters are less broken by combining median blur and dilation, but the characters turn blocky and hence unreadable.

Now contours are found, which return blocky but noise free words. Using the coordinates of the contours, the corresponding regions on the original image are binarized with a lower subtracting constant. This is done to clear the noise away from the characters and only the minimal amount of noise close to the characters remains.

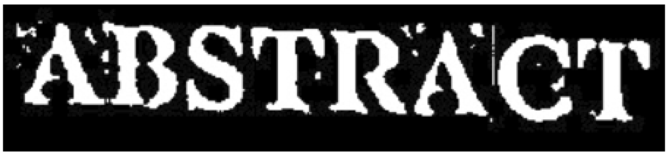


Fig. 8: Binarization on regions of corresponding individual contours

Mask Size = 21 and Subtracting Const = 1

In order to eliminate the noise closest to the characters, we draw contours that surround individual characters and fill the contours with the same color as the background. This is done so that we have an image with only the noise and the words have vanished.

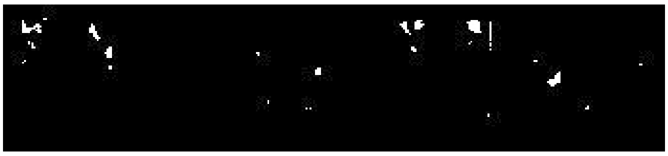


Fig. 9: Noise extraction

Now by subtracting Fig.8 and Fig.9 we would obtain a noise free image ready for character recognition.

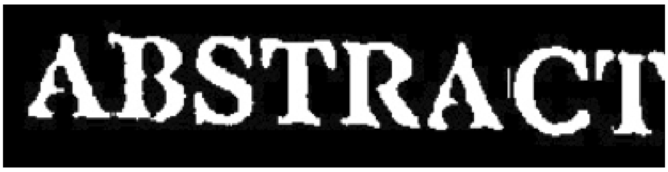


Fig. 10: Noise free image

Notice that Fig.10 is not only noise free, but has better connected characters than Fig.6

3.2 Character Recognition

The noise free image, which is resized to have characters at an average of 20 pixels in height is passed to Tesseract OCR engine to return a string variable of the recognized text.

3.3 Post Processing

3.3.1 Finding Incorrect Words

The string is converted into a list of strings and compared with a dataset containing 466,554 English words. The data set words are present in a SET data structure for a faster lookup table. The approximate time to check if a word exists in the data set or not is 0.000005 sec.

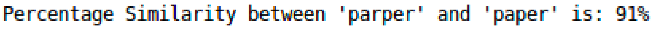
The list of strings are separated into two lists, one for correct and the other for incorrect words.

3.3.2 Checking Recurrences

Since the post processing step is mainly used for further improving the accuracy of an existing highly accurate OCR model, we can assume that the exact same error of recognition may not occur more than once. Thus within the incorrect word list, any word that appears more than once is claimed to be correct. This step helps in correcting nouns such as name of a person, organization, location, which wouldn’t exist in the word data set, and have raised a false alarm.

3.3.3 Similarity comparison of an incorrect word

We use the python package called fuzzywuzzy to compare each word in the incorrect list with the words of the English word data set. Fuzzywuzzy returns a similarity ratio between two words. Hence the incorrect word is corrected to be the word with the highest similarity ratio (the ratio should be above 50%).

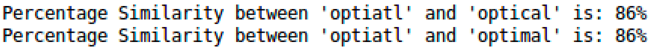


```
Percentage Similarity between 'parper' and 'paper' is: 91%
```

Fig. 11: Fuzzywuzzy Demo

3.3.4 Resolving same similarity percentage

In the event that two or more words from the data set have the same similarity percentage such as Fig.12.



```
Percentage Similarity between 'optiatl' and 'optical' is: 86%
Percentage Similarity between 'optiatl' and 'optimal' is: 86%
```

Fig. 12: Same Similarity Percentage

We use the below probability formulae to pick the best word:

$$probabilty = \frac{Occurrences(Prev_w + Word_i)}{Occurrences(Word_i)}$$

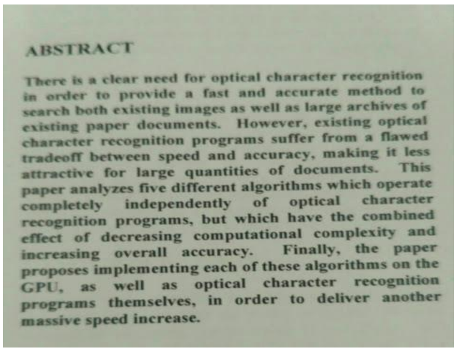
$Prev_w$ - The word before the incorrect word.

$Word_i$ - ith word with same similarity %

By computing the ratio of the number of times the previous of the incorrect word is followed by a candidate word divided by the number of times the candidate word has occurred, we obtain the more probable solution for the incorrect word.

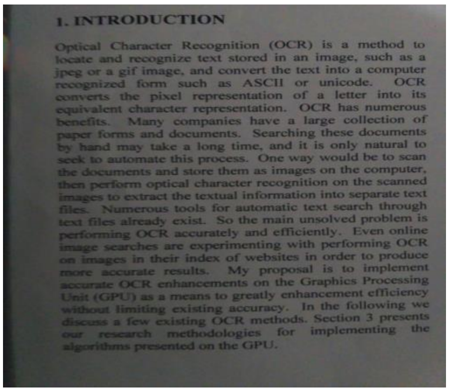
4. Experiment Results

Experiment was conducted on three types of images with poor quality. The poor quality can be classified based on low resolution of the image, low illumination, and high amount of salt and pepper noise.



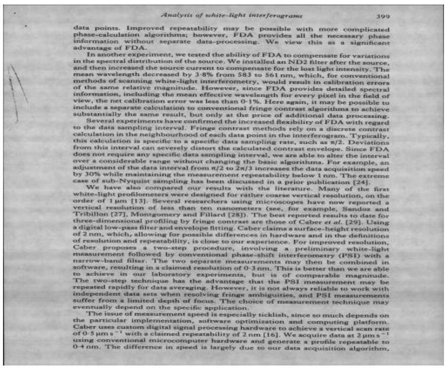
ABSTRACT

There is a clear need for optical character recognition in order to provide a fast and accurate method to search both existing images as well as large archives of existing paper documents. However, existing optical character recognition programs suffer from a flawed tradeoff between speed and accuracy, making it less attractive for large quantities of documents. This paper analyzes five different algorithms which operate completely independently of optical character recognition programs, but which have the combined effect of decreasing computational complexity and increasing overall accuracy. Finally, the paper proposes implementing each of these algorithms on the GPU, as well as optical character recognition programs themselves, in order to deliver another massive speed increase.



1. INTRODUCTION

Optical Character Recognition (OCR) is a method to locate and recognize text stored in an image, such as a jpeg or a gif image, and convert the text into a computer recognized form such as ASCII or unicode. OCR converts the pixel representation of a letter into its equivalent character representation. OCR has numerous benefits. Many companies have a large collection of paper forms and documents. Searching these documents by hand may take a long time, and it is only natural to seek to automate this process. One way would be to scan the documents and store them as images on the computer, then perform optical character recognition on the scanned images to extract the textual information into separate text files. Numerous tools for automatic text search through text files already exist. So the main unsolved problem is performing OCR accurately and efficiently. Even online image searches are experimenting with performing OCR on images in their index of websites in order to produce more accurate results. My proposal is to implement accurate OCR enhancements on the Graphics Processing Unit (GPU) as a means to greatly enhancement efficiency without limiting existing accuracy. In the following we discuss a few existing OCR methods. Section 3 presents our research methodologies for implementing the algorithms presented on the GPU.



Analysis of white-light interferograms 399

data points. Improved repeatability may be possible with more complicated phase-calculation algorithms; however, FDA provides all the necessary phase information without separate data-processing. We view this as a significant advantage of FDA.

In another experiment, we tested the ability of FDA to compensate for variations in the spectral distribution of the source. We installed an ND2 filter after the source, and then increased the source current to compensate for the lost light intensity. The mean wavelength decreased by 3·8% from 583 to 561 nm, which, for conventional methods of scanning white-light interferometry, would result in calibration errors of the same relative magnitude. However, since FDA provides detailed spectral information, including the mean effective wavelength for every pixel in the field of view, the net calibration error was less than 0·1%. Here again, it may be possible to include a separate calculation to conventional fringe contrast algorithms to achieve substantially the same result, but only at the price of additional data processing.

Several experiments have confirmed the increased flexibility of FDA with regard to the data sampling interval. Fringe contrast methods rely on a discrete contrast calculation in the neighbourhood of each data point in the interferogram. Typically, this calculation is specific to a specific data sampling rate, such as $\pi/2$. Deviations from this interval can severely distort the calculated contrast envelope. Since FDA does not require any specific data sampling interval, we are able to alter the interval over a considerable range without changing the basic algorithms. For example, an adjustment of the data interval from $\pi/2$ to $2\pi/3$ increases the data acquisition speed by 30% while maintaining the measurement repeatability below 1 nm. The extreme case of sub-Nyquist sampling has been discussed in a prior publication [24].

We have also compared our results with the literature. Many of the first white-light profilometers were designed for rather coarse vertical resolution, on the order of 1 µm [13]. Several researchers using microscopes have now reported a vertical resolution of less than ten nanometers (see, for example, Sandoz and Tribillon [27], Montgomery and Fillard [28]). The best reported results to date for three-dimensional profiling by fringe contrast are those of Caber *et al.* [29]. Using a digital low-pass filter and envelope fitting. Caber claims a surface-height resolution of 2 nm, which, allowing for possible differences in hardware and in the definitions of resolution and repeatability, is close to our experience. For improved resolution, Caber proposes a two-step procedure, involving a preliminary white-light measurement followed by conventional phase-shift interferometry (PSI) with a narrow-band filter. The two separate measurements may then be combined in software, resulting in a claimed resolution of 0·3 nm. This is better than we are able to achieve in our laboratory experiments, but is of comparable magnitude. The two-step technique has the advantage that the PSI measurement may be repeated rapidly for data averaging. However, it is not always reliable to work with independent data sets when resolving fringe ambiguities, and PSI measurements suffer from a limited depth of focus. The choice of measurement technique may eventually depend on the specific application.

The issue of measurement speed is especially ticklish, since so much depends on the particular implementation, software optimization and computing platform. Caber uses custom digital signal processing hardware to achieve a vertical scan rate of 0·5 $\mu m\ s^{-1}$ with a claimed repeatability of 2 nm [16]. We acquire data at 2 $\mu m\ s^{-1}$ using conventional microcomputer hardware and generate a profile repeatable to 0·4 nm. The difference in speed is largely due to our data acquisition algorithm.

Fig. 13: a) Text Height average 12 pixel b) Poor Illumination c) High Noise

Type of Image	Number of words	Time Taken	Accuracy
Low Resolution	114	12sec	97.4%
Low Illumination	217	20sec	95.3%
High Noise	597	60sec	96.8%

5. Conclusion

This paper has presented an overview of various techniques involving OCR. Though it is not an atomic process it comprises various phases of recognition such as acquiring, pre-processing, classification and post-processing.

We demonstrate how appropriate pre-processing and post processing techniques are required for OCR on digital camera acquired images.

Despite the vast researching in the field of Optical Character Recognition, there are various challenges that still exist such as recognition of characters in various languages, real-time recognition etc. Finally, the use of OCR in real world applications remains an active area of research.

Acknowledgement

We would like to thank our project guide Asst. Proff. K.Deeba, who has been an inspiration for the entire duration of the research. We are also grateful to the Head of the Department Dr. B. Amutha for giving us this opportunity.

References

[1] Image Text To Speech Conversion In The Desired Language By Translating With Raspberry Pi Rithika.H , B. Nithyasanthoshi, IEEE 2016

[2] Image Preprocessing for Improving OCR Accuracy By WojciechBieniecki, Szymon Grabowski and WojciechRozenberg, IEEE July 2007

[3] Real-Time Scene Text Localization and Recognition, Lukáš Neumann, Jiří Matas IEEE 2012

[4] Text Detection and Recognition in Imagery: A Survey Qixiang Ye, Member, IEEE and David Doermann, Fellow, IEEE, published July 2015

[5] Scene text recognition with high performance CNN classifier and efficient word inference XinhaoLiu,TakahitoKawanishi,XiaomengWu,KunioKashino IEEE 2016

[6] Video Text Extraction and Recognition: A Survey, Pooja, RenuDhir IEEE 2016

[7] https://github.com/tesseract-ocr

[8] https://en.wikipedia.org/wiki/Optical_character_recognition

[9] https://pypi.python.org/pypi/pytesseract

[10] Digital Image Processing Book By Rafael C. Gonzalez and Richard Eugene Woods.

[11] T. Padmapriya and V. Saminadan, “Inter-cell Load Balancing technique for multi-class traffic in MIMO-LTE-A Networks”, International Journal of Electrical, Electronics and Data Communication (IJEEDC), ISSN: 2320- 2084, vol.3, no.8, pp. 22-26, Aug 2015.

[12] S.V.Manikanthan and T.Padmapriya “Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g”, International Journal of Pure and Applied Mathematics, ISSN NO:1314-3395, Vol-115, Issue -8, Sep 2017.