Layout Analysis for Scanned PDF and Transformation to the Structured PDF Suitable for Vocalization and Navigation

Azaedeh Nazemi¹, Iain Murray & David A. McMeekin²

¹ Departent of Electrical and Computer Engineering, Curtin University, Perth, WA, Australia

² Department of Spatial Sciences, Curtin University, Perth, WA, Australia

Correspondence: Azaedeh Nazemi, Departent of Electrical and Computer Engineering, Curtin University, Perth, WA, Australia. E-mail: Azadeh.nazemi@postgrad.curtin.edu.au

Received: December 1, 2013	Accepted: December 23, 2013	Online Published: January 19, 2014
doi:10.5539/cis.v7n1p162	URL: http://dx.doi.org/10.3	5539/cis.v7n1p162

Abstract

Information can include text, pictures and signatures that can be scanned into a document format, such as the Portable Document Format (PDF), and easily emailed to recipients around the world. Upon the document's arrival, the receiver can open and view it using a vast array of different PDF viewing applications such as Adobe Reader and Apple Preview. Hence, today the use of the PDF has become pervasive. Since the scanned PDF is an image format, it is inaccessible to assistive technologies such as a screen reader. Therefore, the retrieval of the information needs Optical Character Recognition (OCR). The OCR software scans the scanned PDF file and through text extraction generates an editable text formatted document. This text document can then be edited, formatted, searched and indexed as well as translated or converted to speech. A problem that the OCR software does not solve is the accurate regeneration of the full text layout. This paper presents a technology that addresses this issue by closely preserving the original textual layout of the scanned PDF using the open source document analysis and OCR system (OCRopus) based on geometric layout and positioning information.

The main issues considered in this research are the preservation of the correct reading order, and the representation of common logical structured elements such as section headings, line breaks, paragraphs, captions, and sidebars, foot-bars, running headers, embedded images, graphics, tables and mathematical expressions.

Keywords: optical character recognition, document layout analysis, assistive technology

1. Introduction

For vision-impaired users, access to electronic documents, including scanned PDF files, has been extremely limited. Their main mode of access is by using a screen reader. Even after a scanned document has been processed with OCR software, what is left is plain text without tags or mark-up specifying primitive components. As this resulting document lacks any tags, it is not navigable by vision-impaired users.

In order for readers to navigate scanned PDF, documents need to be tagged by components identification using PDF layout analysis.

Extracting Hidden Structures from Electronic Documents (XED) is a reverse engineering tool for PDF documents (Rigamonti et al., 2005). XED discovers and extracts the original document layout structure, and generates the XCDF hierarchical standard form, which is independent of the document type.

Firstly, XED cleans the primitives in the original document, taking into account all types of embedded resources such as raw images and fonts. Then it recovers the physical structures and represents them in XCDF format. XCDF is able to represent the reorganized document in a structured and unique manner that enables the document content to be accessed easily for further work; however, it is a closed application that works only under the Windows Operating System.

The final target of this research project is to design portable, stand alone, affordable, modifiable and open-source Complete Reading System (CRS) for vision-impaired people (Nazemi & Murray, 2012). CRS provides access to several electronic documents such as Digital Accessible Information System (DAISY), DOC, DOCX, ODT, PDF and all PDF non-textual components such as mathematical expressions and charts. This paper demonstrates part of the development of this open source application that extracts primitive components from an image document,

generates tags, and reconstructs a new document that is accessible and navigable by assistive technologies. The open-source package used for this purpose includes OCRopus for OCR (Shafait, 2009) and ImageMagick for image processing (Still, 2006). The final output is in html Optical Character Recognition (hOCR) format.

2. Layout Analysis

The physical layout analysis represents a document page comprised of unique areas such as columns, paragraphs and text lines (O'Brein, 2012). This layout analysis is responsible for identifying page components such as, text columns, text blocks, text lines and reading order (Breuel, 2008).

Layout is a collection of segments:

$$L = \{S1, ..., Sn\},$$
 where L and S represent layout and segment respectively (1)

A segment is a pixel collection encapsulated within a bounding box defined by its lower left and upper right corner pixels:

$$S = (P1, P2)$$
, where S and P represent segments and pixels accordingly (2)

Each pixel is defined by a coordinate pair:

$$\mathbf{P} = (\mathbf{x}, \mathbf{y}) \tag{3}$$

The layout information is divided into two categories: the geometric layout and the logical layout (Haralick, 1994). The geometric layout is determined by the positioning information about segments. The geometric layout information allows the segments to be categorized into different logical layouts. Each segment is represented with a tag collection. Based on this structure, primary layout analysis is obtained through the following steps:

- computing the bounding box for the connected components of the scanned input page image;
- identifying the whitespace;
- finding the constrained text line.

3. Html Optical Character Recognition (hOCR)

hOCR is a logical format for representing the output of OCR systems. It is an open standard used to embed layout, recognition confidence, style and other information into a recognized text. To successfully embed this data into the text, standard HTML is used. The logical mark-up available in hOCR is designed for the document logical hierarchy, independent of where or how it is rendered on the page. This kind of mark-up is usable for individual documents such as memos and articles, and for compound documents such as newspapers, magazines and collections (Breuel et al, 2007). The hOCR tags that can be used include:

ocr_document, ocr_linear, ocr_title, ocr_author, ocr_abstract, ocr_part, ocr_chapter [H1], ocr_section [H2],

ocr_sub*section [H3, H4], and ocr_par [paragraph].

However, the final output of OCRopus is generated in hOCR format, so it contains only class='ocr_page', class='ocr_par', paragraph separator and a line break
.

This research presents a method for generating hOCR in a fully marked-up format that includes the following additional tags:

ocr_chapter [H1], ocr_section [H2], ocr_sub*section [H3, H4], ocr_image, ocr_table, ocr_math, ocr_caption, ocr_running _footer, OCR _running header, ocr_footbar, and ocr_sidebar.

4. OCRopus Segmentation Methods

Scanned PDF layout analysis strongly depends on the page segmentation method. Recognition by Adaptive Subdivision of Transformation Space (RAST) and Voronoi (named after Georgy Voronoy who created the Voronoi diagram) are two methods used for page segmentation in OCRopus.

RAST extracts connected components and then determines the largest possible whitespace rectangles based on the divider's priority. The RAST algorithm is capable of processing multiple-column documents. In the RAST image result, the column dividers are yellow and different colors are assigned to different segments (Winder, 2010).

The Voronoi method identifies the connected components then extracts sample points along the boundaries to construct a Voronoi-point diagram (Kise et al., 1998). A large number of edges are created, most of which are not required. The unnecessary edges are deleted in an ascending length-wise order, regardless of their connection to other lines. As a result, the Voronoi-point diagram is converted to an area Voronoi diagram, the areas of which represent the page regions.

Moreover, OCRopus provides several tools including:

The ocr-text-image-seg completely separates the image from the text by removing the masked and rectangular regions from an input image.

The ocropus-gpageseg identifies the tops and bottoms of text lines by computing gradients and performing some adaptive thresholding. These components are then used as seeds for the text lines. Ocropus-gpageseg attempts to find column separators as either extended vertical black lines or extended vertical whitespace (OCRopus, 2013).

5. Methodology

This research comprises the following steps for the retrieval of a scanned PDF layout and making it navigable:

- Pre-processing, which includes conversion of image to binary and resizing for segmentation;
- Non-textual extraction of components such as figures and images;
- Block segmentation to divide the page into logical blocks and preserve the reading order. Tables, where applicable, are also detected and extracted at this stage;
- Line segmentation for each text block and computing their lines bounding box in order to recreate the physical layout;
- Geometric data analysis and obtaining the logical layout to generate a tagged document;
- Mathematical expression, detection and extraction;
- Sending the detected tables, figures and/or math expressions to specific applications to extract the valuable and hidden implicit information and represent it as an audio format using Text To Speech(TTS);
- Merging all output components by considering the reading order;
- Generating the hOCR fully marked-up formatted document.

6. Implementation Results

This section describes the implementation results of the present research.

6.1 Non-Text Components Extraction Such as Figures and Images

The ocr-text-image-seg performs document zone classification using run-lengths and connected components based on features and a logistic regression classifier. Since CRS is intended to extract implicit information from non-textual PDF such as figures, another approach is used for figure detection and extraction from image documents which comprises these steps:

- Use RAST segmentation;
- Find all yellow pixels in RAST result;
- Obtain X_{Min} X_{Max} Y_{Min} Y_{Max};
- Check that the pairs of pixels (X_{Min}, Y_{Max})(X_{Max}, Y_{Min}) are located in the yellow area; this means that the yellow area is a rectangle;
- Crop main image from (X_{Min}, Y_{Min}) to (X_{Max}, Y_{Max}) .

The application of RAST for image text separation enables the extraction of the original figure by accessing its bounding box and sending it to GRAPHREADER (Nazemi & Murrey, 2013), which is an application used to extract possible implicit text information from graphical components. After figure extraction from the page has been completed, the figure block is tagged as <ocr-image> </ocr-image> and its bounding box is appended to hOCR.html.

Figure 1 shows a multi-column scanned PDF containing an image, RAST result and extracted figure block. Figure 2 illustrates the result of OCRopus- ocr-text-image-seg.

Digital Talking Book Player for the Visually Impaired Using FPGAs

Azadeh Nazemi, Cesar Ortega-Sanchez, Iair Murray

Electrical and Computer Engineering Depart Curtin University Perth, Western Australia azadeh.nazemi@postgrad.curtin.edu.au, c.ortega@curtin.edu.au, i.murray@curtin.edu.au

Abstrac- The Digital Talking Book (DTB) player is a device for the vis device for the visually impaired to read, search, navigate and bookmark written material using DAISY and EPUB standards. This paper presents the design and implementation of a DTB player in an FPGA-based embedded system to play audio books containing MP3 (Daisy) fills and utilize Text to Speech Synthesis (TTS) for text only books or EPUB books.

I. INTRODUCTION

Visually impaired people traditionally use one of several reading mthods such as Braille and audio tapes. Braille is bulky, has a limited range, is expensive to produce and has a complex coding system (hard to learn) that requires high tactile sensitivity. Analogue books, such as tape and record, or digital books in MP3 format, have sequential access without bookmarking and navigation features [1, 2]. To enable navgation there are two standards: Digital Accessible Information System (DAISY) and Electronic Publication (EPUB), DAISY is an international open standard that (EPOD) PASY is an international open satisfies provides uill navigation and search capabilities, allows bookmarking and can handle human-read and synthetic TTS [3]. EPUE is a text-only standard compatible with the

(3) EPO is a revealed on the second second second and the Daisy standard and has some navigation abilities [4]. Existing DAISY players are either hardware or software based, both of which have significant drawbacks. Hardware players ted to be expensive due to their relatively limited market (compared to mainstream consumer devices). Software ylayers require the user to be familiar with

computers and their often: complex assistive technology. This paper describes an embedded system-on-chip DTB player undr development at Curtin University to address the special realing needs of visually impaired people.

The system is microprocessor-based, containing custom IP, specialzed hardware and a customized Linux kernel. uClinux -an embedded Operating System (OS)- has been installed or MicroBlaze as a suitable platform for hardware /software co-design [5, 6]. The system has been implemented in a Fielt-Programmable Gate Array (FPGA) to allow

reconfiguration of hardware after manufacturing to meet future DTB standards [7, 8]. The final goal is to be able to offer a portable, low-cost,

low-power, fully-accessible digital talking book player for the visually impaired.

II. HIGH-LEVEL SYSTEM DESCRIPTION

Figure 1 shows the high-level block diagram of Curtin's EPUB; uClaux; Text-us-Speech Synthesis
EPUB; uClaux; Text-us-Speech Synthesis
L INTERCENT DESCRIPTION OF A INFORMATION OF A I

with a Personal Computer and a TTS module. The computer supplies the media files to the DTB. Alternatively, a Media Access Controller (MAC) coupled to an on-board physical layer network chp supplies a standard Ethernet connection to get DAISY or EPUB books over an FTP connection. Digilent's PMOD: DA2 and AMP1 [12] are used by an MP3 decoder to play audio files. A user interface based on push huttons com nunicates user requirements and commands to the DTB player. An RC8660 DOUBLETALK module is connected to the

An include of both in the secondary RSIS second prot. This module receives DAISY text files and "reads" them out loud by means of sophisticated text-to-speech algorithms. [13].

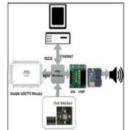


Figure 1. FRGA-based embedded system for DTB player

Digital Talking Book Player for the Visually Impaired Using FPGAs

Azadeh Nazemi, Cesar Ortega-Sanchez, Jain Murrar Electrical and Computer Engineering Depa Curtin University Perth, Western Australia az.deh.nazemi@posterad.curtin.edu.au, c.ortega@curtin.edu.au, i.murrar@curtin.edu.au

Abstract- The Digital Talking Book (DTB) player is a device for the visually impaired to read, search, navigate and bookmark written material using DAISY and EFUB standards. This paper presents the design and implementation of i DTB player in an FPGA-based embedded system tu play audio books containing MP3 (Daisy) files and utilize Text to Speech Synthesis (TTS) for text only books or EPJB books.

Keywords- Assistive Technology; FPGA; Daisy Phyer; EPUB; aClinux; Tent-to-Speech Synthesis

1 INTRODUCTION

Visually impaired people traditionally use one of several reading methods such as Braille and audio tapes, Braille is hulky, has a limited range, is expensive to produce and has a complex coding system (hard to learn) that requires high tactile sensitivity. Anlogue books, such as tape and record, urche ensuriny rimogue come, sect as tape and rector, or digital books in MP3 format, have sequential access without books marking and navigation features [1, 2]. To enable navigation thee are two standards: Digital Accessible Information System (DAISY) and Electronic Publication Information System (DAISY) and Electronic Publication (EPUB): DAISY is an international upen standard that provides full anvigition and search capabilities, allows bookmarking and canhandle human-read and symthet: TTS [3]. EPUB is a tex-only standard compatible with the Daisy standard and has some navigation abilities [4]. Existing DAISY disyste are either hardware or software based, based on which have significant drawbacks. Handware players tend to be expensive due to their relatively limited without comparison of the second standard drawback.

players tento to oc essensive due to men retainvery immed market (compared to mainsteam consumer devices). Software players repoire the user to be familiar with computers and their trees complex assistive technology. This paper describes a embedded system-on-chip DTB player under development at Curtin University to address the

player unour development at Curtin University to address the special reading needs of visually impaired people. The system is microprocessor-based, containing custum IP, specialized hardvare and a customized Linux kentel.

uClinux -an embeddid Operating System (OS)- has been installed un MicroBlaze as a suitable platform for hardware /software co-design [5, 6]. The system has been implemented in a Field-Programmable Gate Array (FPGA) to allow

reconfiguration of hardware after manufacturing to meet

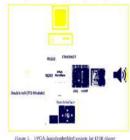
fecontigeration on narrowice area numerosconing, a man-future DTB standards [7, 6]. The final goal is to be able to offer a portable, low-cost, low-power, fully-accessible digital talking book player for the visually impaired.

II. HIGH-LEVEL SYSTEM DESCRIPTION

Figure 1 shows the high-level block diagram of Curtin's DTB player. The current prototype was implemented in the

DTB player. The current prototype was implemented in the SpartandE Starts with board by Digited [10]. The player uses two RS-232 channels the communicate with a Personal Computer and a TTS module. The computer supples the media files to the DTB. Alternatively, a Media Access Controller (MAC) coupled to an on-Social physical layer network chip supples a standard Ethernet connection to get DAISY or EPUB books over an FTP connection. Digiter's PMDD DA2 and AMPL [12] are used by an MP3 decoder to play audio file. A unst interface based on puch home communication area noninnets real connection puch to the network of the standard starts and the start of t aunicates uso requirements and commands to the DTB player. An RC8660 DOUBLETALK module is connected to the

board via a secondary RS32 serial port. This module receives DAISY text files and "reads" them out loud by means of suphisticated tex-to-speech algorithms. [13].



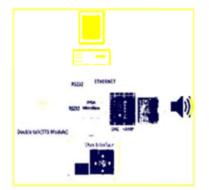


Figure 1. Top left: Multi-column page image containing figure. Top right: RAST image result; Bottom: extracted figure

Digital Talking Book Player for the Visually Impaired Using FPGAs

Azadeh Nazemi, Cesar Ortega-Sanchez, lain Murray AZaUGH Nazethi, Cesa Oriega-Janarica, ani tooriy Electrical and Computer Engineering Department Curtin University Perth, Western Australia azadeh.nazemi@postgrad.curtin.edu.au, c.ortega@curtin.edu.au, i.murray@curtin.edu.au

Abstract— The Digital Talking Book (DTB) player is a rec device for the visually inpaired to read, search, navigate and bookmark written material using DAISY and EPUB standards. This paper presents the design and implementation of a DTB player in an FPGA-based embedded system to play audio books containing MP3 (Daisy) files and uEUT to Speech Synthesis (TTS) for text to show the standard standard standard standard Keywhot Audition Text to Speech Synthesis (TTS) for EVDB, uClinux; Text-to-Speech Synthesis Spite Synthesis (TS) for text to Spite Synthesis (T

I. INTRODUCTION

EPUB; eClinix; Text-o-Speech Synbusy: I. INTCODUCTION Visually impaired people traditionally use one of several fragmethods such as Braille and audio tapes. Braille is public to the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation or digital books in MPJ format, have sequential access of the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation of the sevenation for the sevenation of the sevenation of the sevenation of the sevenation for the sevenation sevenation of the sevenation

reconfiguration of hardware after manufacturing to meet future DTB standards [7, 8]. The final goal is to be able to offer a portable, low-cost, low-power, fully-accessible digital talking book player for the visually impaired.

the visually impaired. II. HIGH-LEVEL SYSTEM DESCRIPTION Figure 1 shows the high-level block diagram of Curtin's DTB player. The current prototype was implemented in the **Sparina3E** Starter kit board by Diglitent [10]. The player uses two RS-232 channels to communicate with a Personal Computer and a TTS module. The computer supplies the media files to the DTB. Altornatively, a Media Access Controller (MAC) coupled to an on-board physical layer network chip supplies a standard Ethernet connection. Diglient's PMODs DA2 and AMPI [12] are used by an MP3 decoder to play audio files. A user interface based on push buttons communicates user requirements and commands to the DTB player.

the DTB player. An RC8660 DOUBLETALK module is connected to the board via a secondary RS232 serial port. This module receives DAISY text files and "reads" them out loud by means of sophsticated text-to-speech algorithms. [13].

Figure 1. FPGA-based embedded system for DTB player

Figure 2. OCR-text-image-seg result

6.2 Block Segmentation

If the PDF page contains only one column, the OCR result lines will be shown in order, but in cases where PDF includes more than one column, there is no guarantee that the OCR result will produce correct ordering of lines. Figure 3 illustrates this ordering issue in a two-column PDF without block segmentation.

accessible books and sources of information. This	research imple	emented using B	Beagle Board	as hardware	platform.
---	----------------	-----------------	--------------	-------------	-----------

Figure 3. Missing reading order in a two-column image document

Figure 4 shows the ordering issue is solved if block segmentation is performed before text line recognition.

accessible books and sources of information. This research

implemented using Beagle Board as hardware platform.

Figure 4. The original reading order is preserved using segmentation

Block segmentation allows the identification of extended vertical black lines or extended vertical whitespace, by using a combination of morphological operations, convolution or convolution and thresholding.

RAST determines the reading order by finding the columns and the text-lines. Figure 5 shows RAST result for a scanned PDF page, which contains two columns and a side bar.

Areassessment of risks and benefits of dopamine agonists in Parkinson's disease

were all ergot

ms abo to the risk of

for e and then

Introduction

Introduction The use of oral dopamine agonists for the treatment of Benefits of dopamine agonists in the medical Parkinson's disease dates back to the 1970s when the management of Parkinson's disease

ergolinic drug bromocriptine was first introduced. The Entymonotherapy first generation of doparnine agonists were all ergo: 0n the basis of a consistent body of evidence from derivatives and their pharmacological profile differed nationised controlled trials, the dopanine agonists form that of levolopa in several ways. For example, ergot-findwaryophic, prepide, granipale, and derivatives had a longer half-life than levolopa and had appairale" has all ben shown to be effective as derivatives nad a tonger naturate than te concept and have a paparet that a text and a set dopamine receptors. Although more than 50 years have inteficity wales stong in twoncripte, chergoline, passed since the non-ergot agonist apomorphine was first and lands at the here at been namined trials of reported to exert strong antipartinsonian effects.¹¹ most these compauds is early Pathiaont disease.¹⁰ In of the currently used non-ergot dopamine agonists have recent jackbootnolled trials, or pinkeli¹⁴ hare been to be effacious in early copiniciple, resignite, provider painteeld.¹⁵ The symptomatic efficacy of dopamine agonists to treat assumption at early used of these drugs as studies have also shown that early use of these drugs as initial monotherapy is associated with a reduced long team inviking tradicative fields in and dyskinesia) compared with levodopa.¹⁵ Although this studies have also to pamine agonists being classified as first-line options for initial monotherapy in early with drainfed related singe uppends of uppends of uppends of uppends of uppends of uppends of the studies have also shown that and intermational granitations compared with levodopa.¹⁵ Although this studies have also as a single discussion as assessed first-line options for initial monotherapy in early with drainfed related single and studies and any mational and intermational granitations of the store pairs for granitational studies the store and the store been teached and provide the store and the associated with levodop in particular store and the store and the

Parkinson's disease in many naconal and international guidelines," there have also been recent concerns about the safety profile of these drugs in the longer term. These dontwost dopamines/ctimulation nd reduced ris concerns are related to the risk of developing impulse of notor complications control discorders, peripheral ocdema, daytime sormolonece. The cut it rason win initial monderapy with a and heart valve fibrosis. The recognition of the risk of dopamine gonistic associate with a reduct risk for and near valve invosts. The reognition of the risk or upunit gives basicities will a different first of a cardiac fibrotic valvulopathies with pergolide and motoromatizations, in particular different academent of the set of the set

ptions with specialised cardiac safety monitoring. more continuous stimulation of bean dopaminergy. In this Review, we first outline the benefits of using receptors and that such continuous stimulation is key to In this Review, we first outline the benefits of using receptor and hat such continuous similation as try to an advance of the second s

cause development of levodopa-induced dyskinesias.* 35 This concept is supported by data from sweral preclinical studies in primates treated with MPTP (I-methyl4

ancet com/neurology Vol 8 October 2009

Figure 5. RAST result segmentation

Pixel colors and position information of RAST result divides the page into blocks according to the following steps:

Eliminate all black (#000000), white (#FFFFFF) and yellow (#FFFF00) pixels; •

167

A reassessment of risks and benefits of dopamine agonists in Parkinson's disease

eral che es of drugs that have been shown to be effective for the treatm have se eat of the recursogens are server concess or origin can have been moving to be enercine to the treatment of the symplomics Parkingson's discuss. Among the first options are the dopamine agonists, which are commonly used both as an early mountherapy and as an adjunct therapy to lendopa. However, before starting any treatment, the overall benefit-orisk rais to individual patients must be considered. For the dopamine agonists, the available evidence on their symplomatic effect, effect to long-term lendopare-table motor complications, patient effect on progression of disease, and adverse event profile must be taken into account. Recendy, the accurrance of adverse events such as leg inu sommolence, impulse control disorders, and fibronis have increasingly ally serious adverse events must therefore be taken into account and t a of viele un affite for individual a

Revie

of 2009; 8: 929-37 Published Online August 25, 2009 DOI:10.1016/51474 4422(09)70225-X Parkinson Institute Clinici di Perfeziona Milan, Italy (A Anto

- Consider color histogram of image;
- Separate pixels based on their color;
- Find the lower left and the upper right corners of pixels for each color;
- Crop the main image from upper right to lower left corners for each color set.

Over-segmentation occurs during Voronoi segmentation. Figure 6 shows over-segmentation in some parts such as in the last paragraph.

III. PREPARE YOUR PAPER BEFORE STYLING

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for $\operatorname{return} f(x) = c_{1} \pm \sum_{n=1}^{\infty} \left(c_{n} \cos^{\frac{n\pi x}{n}} + b_{n} \sin^{\frac{n\pi x}{n}} \right)$

$$you f(x) = a_0 + \sum_{n=1}^{L} \left(a_n \cos \frac{1}{L} + b_n \sin \frac{1}{L} \right)$$

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

Figure 6. Voronoi segmentation result

6.3. Line segmentation

When block segmentation is completed, ocropus-gpageseg runs for each single block to obtain the line segmentation and bounding box of each line. The bounding box recognises position and level number of headings.

Listed below are the abbreviations used for the features extracted from line segmentation and bounding boxes, assuming that each component bounding box is represented by a pair of coordinate values $(x_0, y_0)(x_1, y_1)$ or (x_0, y_0, x_1, y_1) .

line intend or $lm = Left Margin = x_0$ (4)

$$h=y_1 - y_0 \text{ or } h=\$(identify-format "\%h" line.bin.png)$$
 (5)

$$w = x_1 - x_0 \text{ or } w = \$(identify - format ``%w'' line.bin.png)$$
(6)

 $rm = Right Margin = w - x_1 \tag{7}$

ws=White Spaces	(8)
-----------------	-----

$$vs = Vertical Spaces Between Lines$$
 (9)

- no of recognized character=(cat line.txt|sed 's///g'|sed 's/!//g'|wc -c) (10)
- crr = Character Recognition Ratio=number of recognized character /w (11)

ar = aspect - ratio = w/h (12)

$$mean (vs) = l/n(\sum_{i=1}^{n} y_{0_{n}} - y_{1_{n-1}})$$
(13)

ws = (convert line.bin.png line-pixel.txt|cat line-pixel.txt|grep - c FFFFFF) (14)

based on number of white pixels in line segment

$$ws_{mean} = (ws_n + ws_{n+1})/2$$
 (15)

$$ratio(lm) = x_0 / min(lm)$$
(16)

$$ratio(ws) = 2ws_n / ws_{n+} ws_{n+1} \tag{17}$$

$$h_{mean} = l/n \sum_{i=1}^{n} h_i \tag{18}$$

$$ratio(h) = h/mean(h)$$
(19)

$$ratio(aspect) = aspect/mean(aspect)$$
(20)

Based on the extracted features from the bounding box of each single line, the following can be concluded:

- A position is reserved as a heading level if: hOCR file indicates the position as a new paragraph $\langle P \rangle$; $x_0 > minimum(lm)$; $y_{1-}y_{0>}mean(h)$
- Caption tag is assigned to a line segment if: $w=x_0 + x_1$ (w is the width of line segment) and the previous segment is an image or the next segment is a table
- Running header tag is attached to a line segment if it is the first line segment of the first block and vertical space between this line and the next line >mean value (vertical spaces in page); y_{1-y0>}mean(h)
- Running footer tag is attached to a line segment if it is the last line segment of the last block and vertical space between this line and the previous line >mean value (vertical spaces in page); $y_1 y_{0} = mean(h)$
- Side bar tag is assigned to the line segment if it is located in the last block in a page and the block-aspect-ratio<1
- Foot-bar tag is assigned to the line segment if it is located in the last block in a page and the block-aspect-ratio>1

6.4 Table Recognition

When applying the RAST segmentation method to a document image that contains a table, the horizontal and vertical separator lines for table cells are specified in yellow. The coordinate values of line intersections, which are in fact the bounding box of each cell, are obtained by computing the number of these lines and finding their geometric properties. Therefore, by using these bounding box values, all cells can be separated from the table image as individual segments and marked with tags $\langle ocr - table - cell_{ij} \rangle \langle /ocr - table - cell_{ij} \rangle$, where i and j represent row and column respectively. Figure 7 illustrates a sample table and its RAST output.

Community Courses Bath Automa 1997				Courses Beth Autom 1997					
Course Name	Course Tutor	r Summary	Code	Fee	Course Name	Course Tutor	Summery		Fee
After the Civil Wer	Dr. John Wroughton	The course will examine the turbulent years in Englani after 1646. 6 weekly meetings starting Monday 13th October.	H27	£32	After the Civil Wer	Dr. John	The course will exemine the turbulent years in England after 1646. 6 weakly meetings starting Monday 13th October.	H27	£ 32
An Introduction to Anglo-Sexon England	Mark Cottle	One day course introducing the early medieval period reconstruction the Angle-Saxons and their society. Saturday 18th October.	H28	£18	An Introduction to Anglo-Saxon England	Mark Cottle	One day course introducing the early medieval period reconstruction the Anglo-Sexons and their society. Sanaday 18th October.	H28	£18
The Glory that was Greece	Velezie Lorenz	Birthplace of denocracy, philosophy, heardand of theater, home of argument. The Romans may have done it but the Greeks did it first. Saturday day school 25th October 1997	H30	£18	The Glory that was Greece	Velerie Lorenz	Birthpiace of democracy, philosophy, heardand of theates, home of argument. The Romans may have done it but the Oreeks didit fust. Saturday day school 25th October 1997	H30	£18

Figure 7. Sample of table and its RAST result

7. Conclusion and Further Works

This research is intended to generate a complete hOCR file from the results of OCRopus by adding appropriate tags. These tags are utilised to retain the reading order, detect table cells and navigate by row or column, specifying various heading levels within the document image.

The application developed from this research sends a detailed hOCR file to an intermediate module. To provide navigation ability, this module is responsible for distinguishing all tags and extracting individual text parts of the document based on user requests.

The remaining issues regarding this subject include mathematical expression detection, extraction and recognition within image documents. There are several attributes, which must be considered when dealing with image documents containing mathematical expressions such as: multi-dimensionality, lack of ordering compared to plain text and over-segmentation during line segmentation. Although some research has already been conducted in an attempt to address these issues (Garain, 2009; Yamazaki et al., 2011), there is a need for further development in order to design a comprehensive mathematical detection system by using these methods to make them accessible by assistive technology for vision-impaired users.

References

- Breuel, T. M. (2007, September). The hOCR microformat for OCR workflow and results. In *Document Analysis* and Recognition, 2007. ICDAR 2007. Ninth International Conference on (Vol. 2, pp. 1063-1067). IEEE.
- Breuel, T. M. (2008). The OCRopus open source OCR system. In Proc. SPIE Document Recognition and Retrieval XV (pp. 0F1-0F15). San Jose, CA, USA.
- Garain, U. (2009, July). Identification of mathematical expressions in document images. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 1340-1344). IEEE. http://dx.doi.org/10.1109/ICDAR.2009.203
- Haralick, R. M. (1994, June). Document image understanding: Geometric and logical layout. In Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on (pp. 385-390). IEEE. http://dx.doi.org/10.1109/CVPR.1994.323855
- Kise, K., Sato, A., & Iwata, M. (1998). Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding*, 70(3), 370-382. http://dx.doi.org/10.1006/cviu.1998.0684
- Nazemi, A., & Murray, I. (2012). A Novel Complete Reading Embedded System for the Vision Impaired. 3rd Annual International Conference on Computer Science Education Innovation and Technology. Singapore. http://dx.doi.org/10.5176/2251-2195 CSEIT12.40#sthash.6PhwdNiv.dpuf
- Nazemi, A., & Murray, I. (2013). A Method to Provide Accessibility for Visual Components to Vision Impaired. International Journal of Human Computer Interaction (IJHCI), 4(1), 54.
- O'Brien, S. (2012). Optical Character Recognition. Worcester Polytechnic Institute.
- Rigamonti, M., Bloechle, J. L., Hadjar, K., Lalanne, D., & Ingold, R. (2005). Towards a canonical and structured representation of PDF documents through reverse engineering. Paper presented at the *Document Analysis* and Recognition, 2005. Proceedings. Eighth International Conference on., Vol. 2 (pp. 1050-1054).
- Shafait, F. (2009). Document image analysis with OCRopus. Paper presented at the *Multitopic Conference*, 2009. *INMIC 2009. IEEE 13th International.*, 1(6), 14-15. http://dx.doi.org/10.1109/INMIC.2009.5383078

Still, M. (2006). The Definitive Guide to ImageMagick.

The OCRopus(tm) open source document analysis and OCR system (April, 2013). Retrieved from https://code.google.com/p/ocropus

Winder, A. A. (2010). Extending the Page Segmentation Algorithms of the OCRopus Documentation Layout Analysis System (Doctoral dissertation, Boise State University).

Yamazaki, S., Furukori, F., Zhao, Q., Shirai, K., & Okamoto, M. (2011, September). Embedding a Mathematical OCR Module into OCRopus. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (pp. 880-884). IEEE. http://dx.doi.org/10.1109/ICDAR.2011.180

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).