Improving Digital Library Support for Historic Newspaper Collections

A thesis
submitted in partial fulfilment
of the requirements of the degree
of

Master of Science in Computer Science

at

The University of Waikato

By

Leo Lin

The University of Waikato
2009

Abstract

National and international initiatives are underway around the globe to digitise the vast treasure troves of historical artefacts they contain and make them available as digital libraries (DLs). The developed DLs are often constructed from facsimile pages with pre-existing metadata, such as historic newspapers stored on microfiche or generated from the non-destructive scanning of precious manuscripts. Access to the source documents is therefore limited to methods constructed from the metadata. Other projects look to introduce full-text indexing through the application of off-the-shelf commercial Optical Character Recognition (OCR) software. While this has greater potential for the end user experience over the metadata-only versions, the approach currently taken is "best effort" in the time available rather than a process informed by detailed analysis of the issues. In this thesis, we investigate if a richer level of support and service can be achieved by more closely integrating image processing techniques with DL software.

The thesis presents a variety of experiments, implemented within the recently published open-source OCR System (*Ocropus*). In particular, existing segmentation algorithms are compared against our own based on Hough Transform, using our own created corpus gathered from different major online digital historic newspaper archives.

Acknowledgements

There are many people to be acknowledged through the progress of this thesis. My sincere appreciation to my supervisor Dr. David Bainbridge, for his patient guidance throughout the research and experiments for this thesis, and the innovative idea suggested, without his efforts this thesis could not have been accomplished.

My acknowledgement to the members of the *Ocropus* group especially Thomas, Faisal, Ilya and Christian. For your hints and suggestions through the implementation of algorithms, without your contribution the experiments would not have progressed as smoothly.

My thanks to Mrs. Jenny McLarens for proofreading this thesis, without you it would have been tedious. Finally, with my gratitude to my family (Sam, Gloria, Jamie), friends and members of Waikato Kendo Club, thank you all for the great support and encouragement, throughout the most challenging time of my study. Thank you for being there for me.

Leo Lin

University of Waikato

April 2008

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

Libraries and archives are increasingly digitising physical resources and presenting the result of such labours as digital libraries (DLs), available for the world to access. Collections, such as historic newspapers, are a particularly popular choice to digitise as they are usually free of copyright restrictions. A common practice for making an historical newspaper searchable is to apply off-the-shelf commercial OCR (Optical Character Recognition). Given the budget and time-constraints of such projects, it is fair to say the procedure has been less than scientific.

Furthermore, the typography of historic newspapers presents several novels and interesting challenges, which have not yet been taken into account by present work (both research and available software). For that reason the endeavour of this thesis is to conduct a more rigorous evaluation of relevant techniques.

1.1 Background

Historical newspapers often do not have an electronic copy available, and many of them are already out of copy. Even with the original copy available, the accessibility to such a copy is often difficulty. Many of them can only be accessed from the national library of one's country and special authorisation might be required. Pests, humidity and natural disasters potentially pose the greatest threat to their preservation in physical form.

Fortunately the advancement of the technological devices such as microfilm, scanner and digital camera, provide alternative solutions for the preservation of historical newspapers. Particularly with scanner and digital camera capturing scan or image, we can then store the digitised images into the computer's repository, in such way the image itself is protected from deterioration for the time being. Numerous copies can then also be made effortlessly for backup purposes. Technical obsolescence is the main spectre to digital preservation, and strategies to guard against this are well documented (Witten, Bainbridge 2002).

Newspapers, especially historical newspapers hold a significant value to the local community, one's nation and even to the whole world. Historical newspapers can be published daily, weekly or even monthly, and feature articles and columns expanded over the flow of time, recording the histories and events of the communities during a particular time frame of human history, with a substantial amount of detail and information. Such information can become an immense asset not only to the general community, but to all the academic communities in the assistance of the research. In addition to the above it is also a valuable tool to historians, economists, ecologists, geologists, and the like.

Conversely, such historical treasures of human history will become less practical if we do not have the appropriate portal to access all the information efficiently. Blessed by the technology of today, we are now able to digitise all the historical newspapers into computer repositories using a variety of devices. Nonetheless this is only stage one towards storing the treasures, as the digitised newspapers will be saved in various image format (JPEG, PNG, BMP, TIFF, GIF). Therefore the contents cannot be directly accessible. This is to say we cannot simply apply a search to its content like we could in Google (to find relevant web pages).

In order for the content to be searchable, a typical way is to apply OCR to every image that we digitised. (An example will be given in Section 1.3). The technique of OCR here refers to the method of converting each character within a digital image into computer readable text (ASCII or Unicode text). In such a way we can then obtain the searchable text and link it with the digitised image.

After decades of research and development (Mori et al. 1992) the technique of OCR has become reasonably mature, with a high accuracy of recognition rate. Once we have both digitised image and the text after OCR, we can then use these documents to construct a digital library collection.

1.2 Synopsis

The structure of this thesis is as follow. An example of a historic newspaper DL is given in the next section; followed by a brief overview of a modern open source digital library system, Greenstone. In Chapter 2 we review the image-processing and document layout analysis algorithms the work is based upon. Related projects are also summarised. In addition, possible frameworks for implementation of digitised image analysis are introduced.

The software implementation of the chosen framework is explained in Chapter 3 as well as the corpus used for the experiments and the reason for choosing them. The results of the devised are presented experiments in Chapter 4. Different experiments have been run based on different algorithms, in order to better understand and evaluate their performance and characterises when processing historic newspaper images. An experiment has also been performed by using a combination of different algorithms, for the evaluation of the segmentation performance on the historical newspaper layout. Chapter 5 provides a summary and conclusions of the work.

1.3 Walk through of "Papers Past"

In this section demonstrate an example of an historic newspapers collection website call *"Papers Past"* (Papers Past 2009). The website is provided by the *National Library of New Zealand* (NATLIB NZ 2009). It was launched in year 2001 with 250,000 pages of historic New Zealand newspapers.

In the initial version, no text was available for indexing, and the only way to find content was to browse your way to a particular newspaper. It was essentially the digital equivalent of using a microfiche reader.

New pages are added into the collections regularly, and in year 2007 the website was re-launched with new user interface, in addition to the new full-text search functionality based on OCR'd data for one third of the collection. Their goal is to eventually make the whole collection searchable. One of the main reasons for introducing full-text search is based on the response of users of *Papers Past*, as is mentioned in the following statements.

"Papers Past was digitised images, so using it was difficult. There was a lot of feedback from users who wanted better access." says Tracy Powell, Project Leader,

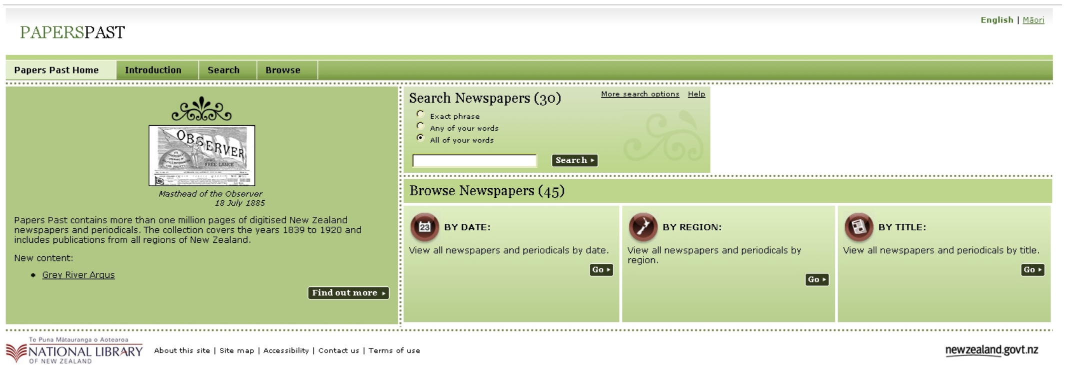
Innovation Centre, NLNZ (Papers Past case study 2008). Undeniably the collection with such a volume will appear to be extremely difficult for any users to locate their desired information.

At the time of writing, *"Papers Past"* holds slightly more than one million digitised New Zealand newspapers from all regions of New Zealand, with collections covering the years from 1839 to 1920.

In terms of the technology used, the re-launched *Papers Past* utilises *Greenstone* (detail in Section 1.4) as its primary collection constructing system. In order to make the collection full-text searchable, use of the off-the-shelf commercial OCR software was introduced for automatically generating searchable newspaper text. The main part of the budget was spent on having the regions of the paper categorised by hand.

The text is then indexed using Greenstone and associated with the corresponding digitised image, though at this stage due to the vast amount of data, the website is yet unable to provide the user with a manually corrected version of text i.e. a lot of mistakes are to be expected.

These mistakes will substantially decrease the amount of information that users can access. For example, if the user search for the word "rugby" in the collection, but a proportion of the text within the collection has been recognised as "rugbv" during the procedure of OCR, these results will be missed.



Search Newspapers (30)

○ Exact phrase
○ Any of your words
● All of your words

Masthead of the Observer
18 July 1885

Papers Past contains more than one million pages of digitised New Zealand newspapers and periodicals. The collection covers the years 1839 to 1920 and includes publications from all regions of New Zealand.

New content:

- Grey River Argus

Browse Newspapers (45)

BY DATE: View all newspapers and periodicals by date.

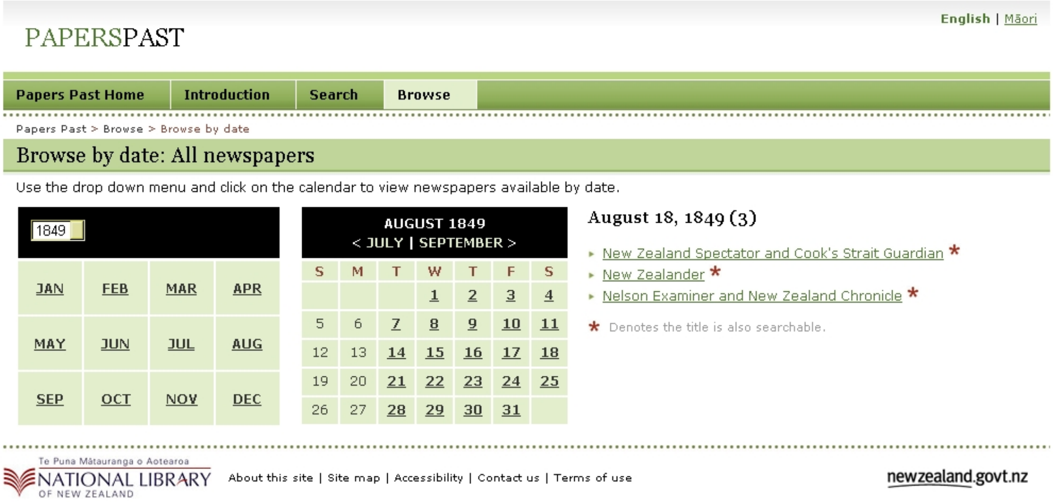
BY REGION: View all newspapers and periodicals by region.

BY TITLE: View all newspapers and periodicals by title.

Figure 1.3a: Papers Past website.

Figure 1.3a illustrates the main website of *Papers Past*. When we first arrive at the site, the system will display one issue of a newspaper by randomly selecting a date and region from the collection. The website provides three different browsing methods for the user: they can browse the collection by date, region or title. In addition the user can simply enter the phrases or any words in the search textbox provided above the three browsing options.

By clicking on the “by date” link on Figure 1.3a, the “browse by date” page will be shown to the user. The user can then select the date in order of year, month followed by day, and a number of newspapers associated with such date will then be revealed. However, not all the months and days can be selected. Those that can be selected are indicated by underlining.


PAPERSPAST
English | Māori
Papers Past Home
Introduction
Search
Browse
Papers Past > Browse > Browse by date
1849
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
< JULY | SEPTEMBER >
Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND
About this site | Site map | Accessibility | Contact us | Terms of use
newzealand.govt.nz

Browse by date: All newspapers

Use the drop down menu and click on the calendar to view newspapers available by date.

AUGUST 1849

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

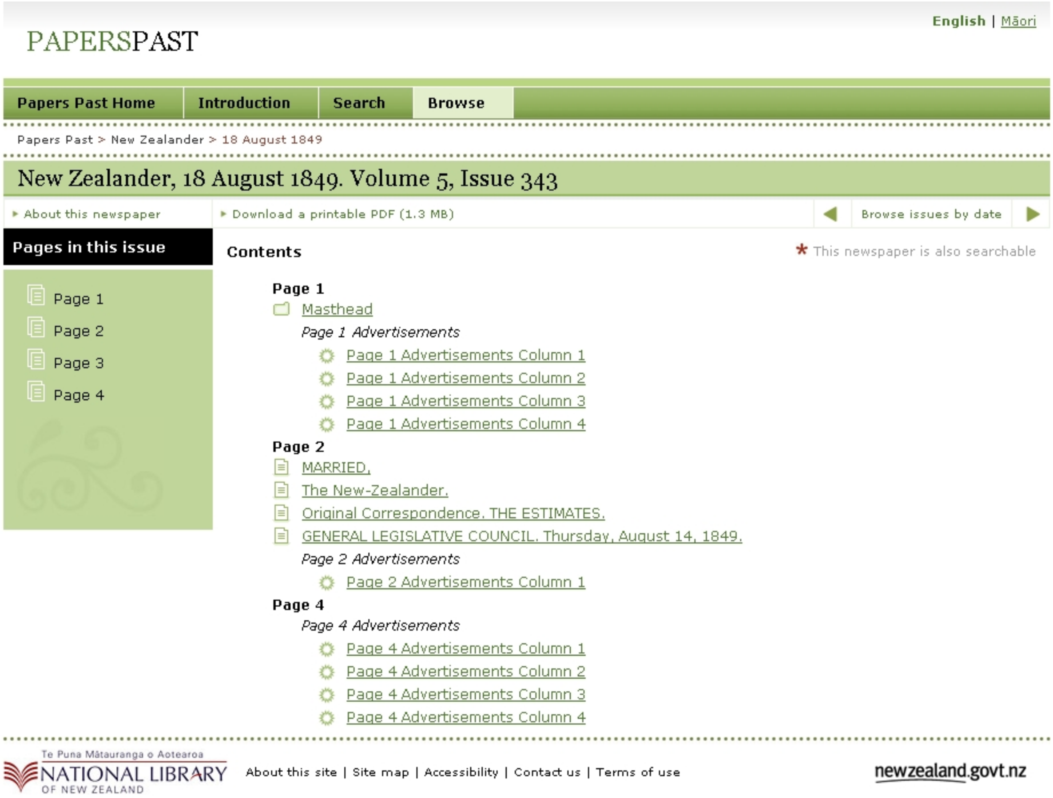
August 18, 1849 (3)

- New Zealand Spectator and Cook's Strait Guardian *
- New Zealander *
- Nelson Examiner and New Zealand Chronicle *

* Denotes the title is also searchable.

Figure 1.3b: Browse by date structure.

As we can see from Figure 1.3b three newspapers are shown to the user, as three of them are all associated with the date August 18, 1849. The red asterisk icon at the end of each newspaper title indicates that the title is also searchable by the user. This is an example of finding a particular item within the collection by searching its title metadata in Greenstone. If the user wishes to view the actual digitised image of one of the newspapers, they can do so by clicking on the title.



PAPERSPAST

English | Māori

Papers Past Home | Introduction | Search | Browse

Papers Past > New Zealander > 18 August 1849

New Zealander, 18 August 1849. Volume 5, Issue 343

▸ About this newspaper

▸ Download a printable PDF (1.3 MB)

◀ Browse issues by date ▶

Pages in this issue

- Page 1
- Page 2
- Page 3
- Page 4

Contents

* This newspaper is also searchable

Page 1

- Masthead

Page 1 Advertisements

- Page 1 Advertisements Column 1
- Page 1 Advertisements Column 2
- Page 1 Advertisements Column 3
- Page 1 Advertisements Column 4

Page 2

- MARRIED.
- The New-Zealander.
- Original Correspondence. THE ESTIMATES.
- GENERAL LEGISLATIVE COUNCIL. Thursday, August 14, 1849.

Page 2 Advertisements

- Page 2 Advertisements Column 1

Page 4

Page 4 Advertisements

- Page 4 Advertisements Column 1
- Page 4 Advertisements Column 2
- Page 4 Advertisements Column 3
- Page 4 Advertisements Column 4

Te Puna Mātauranga o Aotearoa NATIONAL LIBRARY OF NEW ZEALAND

About this site | Site map | Accessibility | Contact us | Terms of use

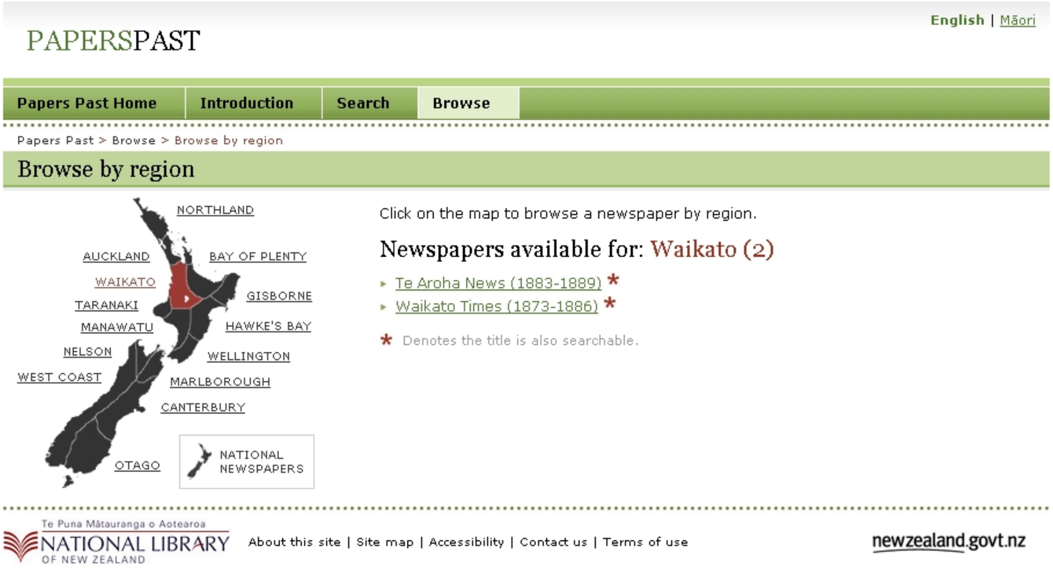
newzealand.govt.nz

Figure 1.3c: Details of one newspaper issue.

By clicking on one of the displayed newspaper titles *"New Zealander"*, we will then be shown the layout information of this particular issue. As we can see from Figure 1.3c it specifies on the left hand side the number of pages in this issue, of which only four pages in this case, can be directly accessed.

In addition, the content of each page is listed on the right hand side. The content structure for each page is primarily based on columns, which is very similar to the structure of contemporary newspapers. As in Figure 1.3c there are four columns identified under Page one. If the user wishes to access one particular column as an alternative to the whole page, they can do so by clicking on the column link. In some cases the content has been enlarged with supplementary details, such as in Page Two, four subtitles have been listed.

These are not based on the structure of columns, but instead they are based on the title of each article or events. To a certain extent such structure is more user driven as it offers users more information before they access the image. Considerably more logical than column one, column two, column three, et cetera.



PAPERSPAST

English | Māori

Papers Past Home | Introduction | Search | Browse

Papers Past > Browse > Browse by region

Browse by region

Click on the map to browse a newspaper by region.

Newspapers available for: Waikato (2)

- Te Aroha News (1883-1889) *
- Waikato Times (1873-1886) *

* Denotes the title is also searchable.

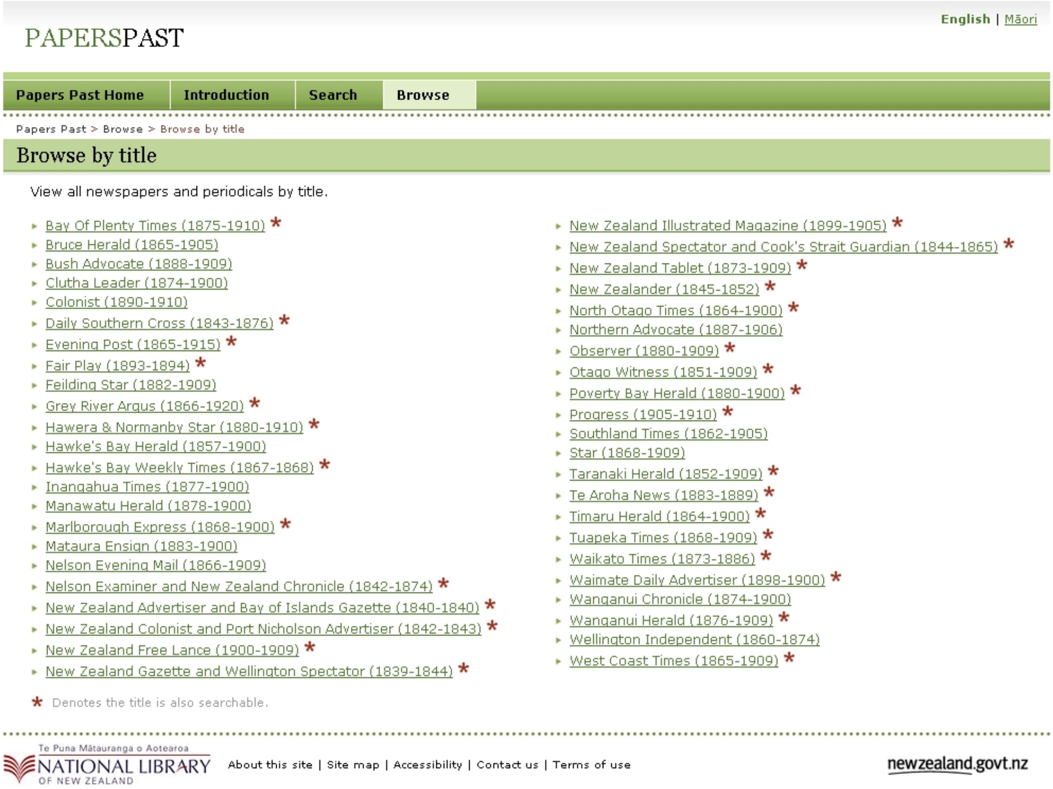
Te Puna Mātauranga o Aotearoa NATIONAL LIBRARY OF NEW ZEALAND

About this site | Site map | Accessibility | Contact us | Terms of use

newzealand.govt.nz

Figure 1.3d: Browse by region structure.

An alternative way to browse newspapers is by its region. By clicking on a region icon from Figure 1.3a, the page with a map of New Zealand will be shown. Clicking the region that we want to view, will access the newspapers available for this region and will then be displayed on the right hand side. In Figure 1.3d we selected the Waikato region. Two newspapers have been listed. Both newspapers list the years of the available issues in the current collection. By clicking on the title of newspaper selected we will then be able to access this particular issue by either browsing, by date, or searching within this newspaper.



Browse by title

View all newspapers and periodicals by title.

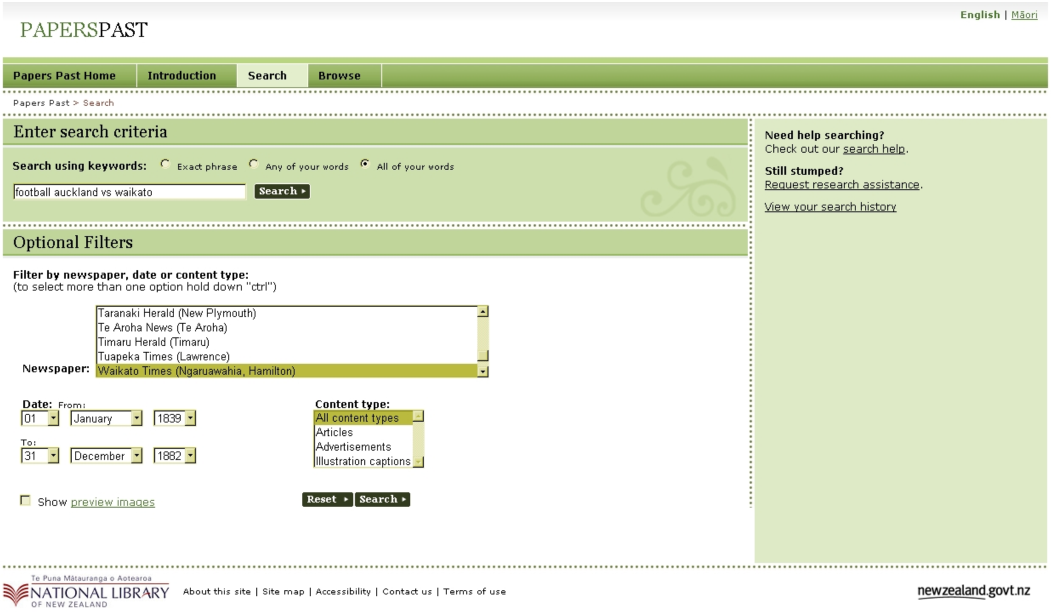
- Bay Of Plenty Times (1875-1910) *
- Bruce Herald (1865-1905)
- Bush Advocate (1888-1909)
- Clutha Leader (1874-1900)
- Colonist (1890-1910)
- Daily Southern Cross (1843-1876) *
- Evening Post (1865-1915) *
- Fair Play (1893-1894) *
- Feilding Star (1882-1909)
- Grey River Argus (1866-1920) *
- Hawera & Normanby Star (1880-1910) *
- Hawke's Bay Herald (1857-1900)
- Hawke's Bay Weekly Times (1867-1868) *
- Inangahua Times (1877-1900)
- Manawatu Herald (1878-1900)
- Marlborough Express (1868-1900) *
- Mataura Ensign (1883-1900)
- Nelson Evening Mail (1866-1909)
- Nelson Examiner and New Zealand Chronicle (1842-1874) *
- New Zealand Advertiser and Bay of Islands Gazette (1840-1840) *
- New Zealand Colonist and Port Nicholson Advertiser (1842-1843) *
- New Zealand Free Lance (1900-1909) *
- New Zealand Gazette and Wellington Spectator (1839-1844) *
- New Zealand Illustrated Magazine (1899-1905) *
- New Zealand Spectator and Cook's Strait Guardian (1844-1865) *
- New Zealand Tablet (1873-1909) *
- New Zealander (1845-1852) *
- North Otago Times (1864-1900) *
- Northern Advocate (1887-1906)
- Observer (1880-1909) *
- Otago Witness (1851-1909) *
- Poverty Bay Herald (1880-1900) *
- Progress (1905-1910) *
- Southland Times (1862-1905)
- Star (1868-1909)
- Taranaki Herald (1852-1909) *
- Te Aroha News (1883-1889) *
- Timaru Herald (1864-1900) *
- Tuapeka Times (1868-1909) *
- Waikato Times (1873-1886) *
- Waimate Daily Advertiser (1898-1900) *
- Wanganui Chronicle (1874-1900)
- Wanganui Herald (1876-1909) *
- Wellington Independent (1860-1874)
- West Coast Times (1865-1909) *

* Denotes the title is also searchable.

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND
About this site | Site map | Accessibility | Contact us | Terms of use
newzealand.govt.nz

Figure 1.3e: Browse by title structure.

The last browsing structure is “browsing by title”, which lists the entire collection of newspapers, along with the years of available issues. As Figure 1.3e shows we can only see a portion of newspapers with red asterisk at the end, which indicates they are searchable. Those without the asterisk at this stage can only be browsed.


PAPERSPAST
English | Māori
Papers Past Home
Introduction
Search
Browse
Papers Past > Search
Search ▸
Reset ▸
Search ▸
Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND
About this site | Site map | Accessibility | Contact us | Terms of use
newzealand.govt.nz

Enter search criteria

Search using keywords: ○ Exact phrase ○ Any of your words ● All of your words

football auckland vs waikato

Optional Filters

Filter by newspaper, date or content type:
(to select more than one option hold down "ctrl")

Newspaper:
Taranaki Herald (New Plymouth)
Te Aroha News (Te Aroha)
Timaru Herald (Timaru)
Tuapeka Times (Lawrence)
Waikato Times (Ngaruawahia, Hamilton)

Date: From: 01 January 1839
To: 31 December 1882

Content type:
All content types
Articles
Advertisements
Illustration captions

□ Show preview images

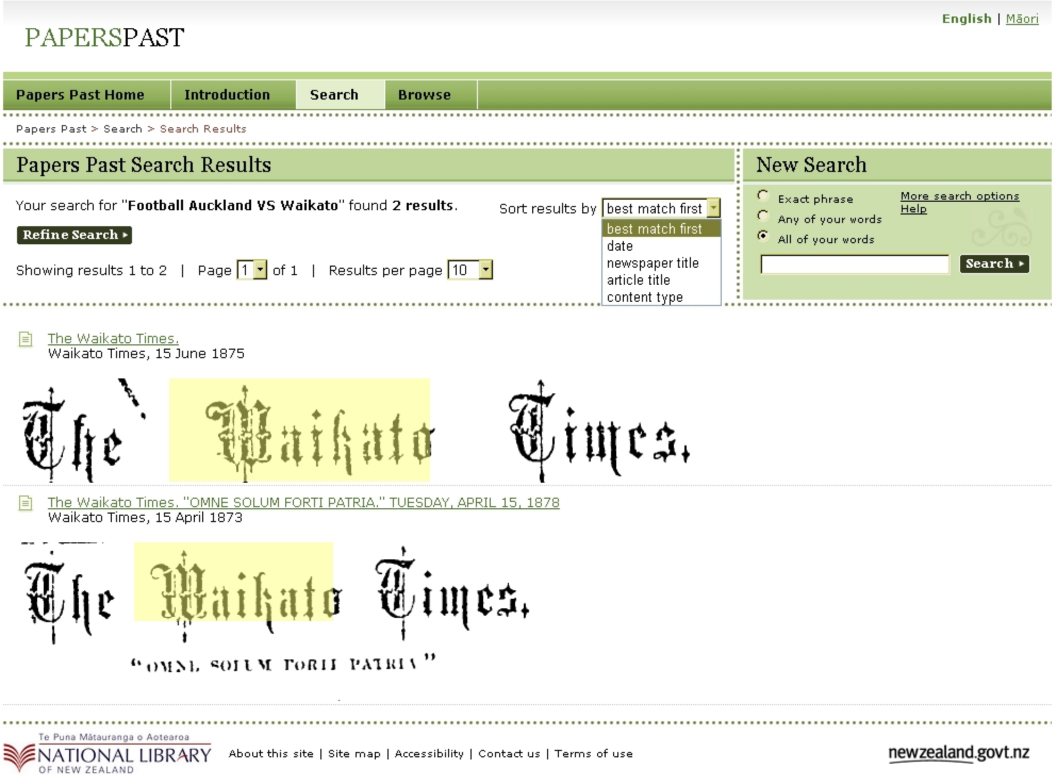
Need help searching?
Check out our search help.

Still stumped?
Request research assistance.

View your search history

Figure 1.3f: Search page with optional filters.

Recently the search function has been introduced to *Papers Past* website: it increases the user's accessibility to related newspapers. We can access the search function by clicking on to the third button in the menu bar. As in Figure 1.3f we have three options for search keywords. We can choose from "exact phrase", "Any of your words", or "All of your words", and the results will be based on the option we choose. Underneath the search textbox optional filters are provided for the user, which allows them to narrow down the newspaper (region indicated), date or content type and the option to display a preview image with the results.



PAPERSPAST

English | Māori

Papers Past Home | Introduction | Search | Browse

Papers Past > Search > Search Results

Papers Past Search Results

Your search for "**Football Auckland VS Waikato**" found **2 results**.

Refine Search ▸

Showing results 1 to 2 | Page 1 of 1 | Results per page 10

Sort results by best match first

- best match first
- date
- newspaper title
- article title
- content type

New Search

○ Exact phrase
○ Any of your words
● All of your words

More search options
Help

Search ▸

The Waikato Times.
Waikato Times, 15 June 1875

The Waikato Times. "OMNE SOLUM FORTI PATRIA." TUESDAY, APRIL 15, 1878
Waikato Times, 15 April 1873

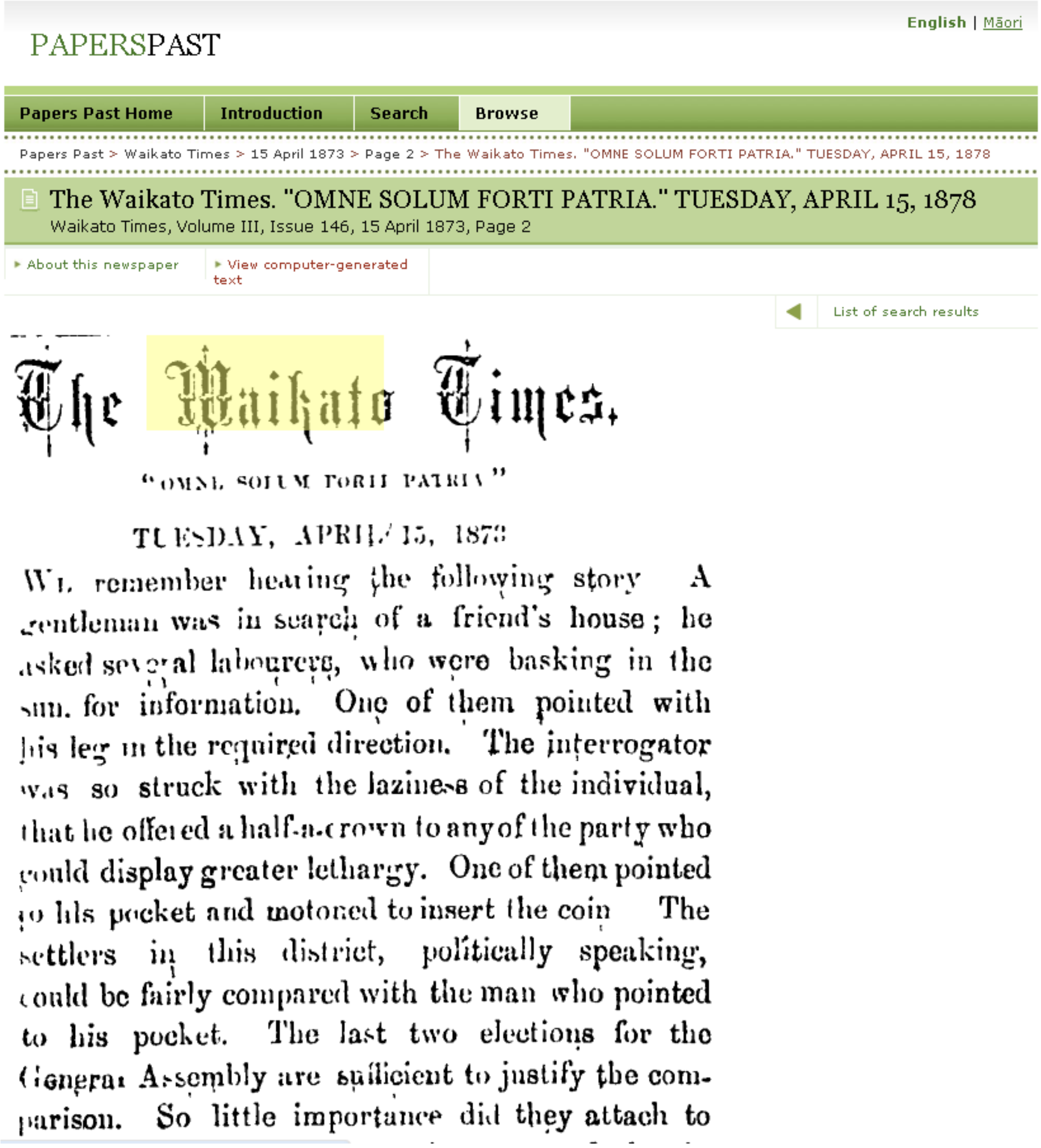
Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

About this site | Site map | Accessibility | Contact us | Terms of use

newzealand.govt.nz

Figure 1.3g: Search results.

Figure 1.3g illustrates the results of our searching, with the keyword *"Football Auckland VS Waikato"*, and with following filter options, Newspaper: Waikato Times, Date: 01 January 1839 – 31 December 1882, Content Type: All content types. Two results have been returned from the collection. In addition the system highlights the word on the newspaper image that matched with our keyword. By clicking on to the desired newspaper title we will able to see the digitised image of the corresponding article. Additionally, the website also provides a sorting function which allows the user to sort the results by either "best match first", "date", "newspaper title", "article title", or "content type".



English | Māori

PAPERSPAST

Papers Past Home | Introduction | Search | Browse

Papers Past > Waikato Times > 15 April 1873 > Page 2 > The Waikato Times. "OMNE SOLUM FORTI PATRIA." TUESDAY, APRIL 15, 1878

The Waikato Times. "OMNE SOLUM FORTI PATRIA." TUESDAY, APRIL 15, 1878

Waikato Times, Volume III, Issue 146, 15 April 1873, Page 2

▸ About this newspaper ▸ View computer-generated text

◀ List of search results

The Waikato Times.

"OMNE SOLUM FORTI PATRIA."

TUESDAY, APRIL 15, 1873

We remember hearing the following story. A gentleman was in search of a friend's house; he asked several labourers, who were basking in the sun, for information. One of them pointed with his leg in the required direction. The interrogator was so struck with the laziness of the individual, that he offered a half-a-crown to any of the party who could display greater lethargy. One of them pointed to his pocket and motioned to insert the coin. The settlers in this district, politically speaking, could be fairly compared with the man who pointed to his pocket. The last two elections for the General Assembly are sufficient to justify the comparison. So little importance did they attach to

Figure 1.3h: Digitised image from one of the newspapers in search results.

Figure 1.3h shows an excerpt of digitised newspaper images from the previous search results. Just above the image there is a red highlighted text "View computer-generated text" which gives the option to view the text generated from off-the-shelf commercial OCR software.

Here is the OCR text up until the first semicolon:

“The Waikato Times. "OMNE SOLUM FORTI PATRIA." TUESDAY, APRIL 15, 1878 Wi, remember lilmi in<^ yhe following story A w 'entleinan was in search of a friend's house ;”

As we can see from the OCR text above, there are a significant amount of recognition errors except for the first line, which lists the name of newspaper, article title and the date, which could be manually inputted by the human editor. There are many factors which will cause recognition error; we will return to this point in Chapter 2.

1.4 Greenstone Digital Library

The past decade has seen Digital Library (DL) software transition from research systems to production level software, deployed by countless organisations around the world. Among others, *Greenstone* is one of the most comprehensive and mature open-source software systems (Witten et al. 2000). The system allows the users to gather different types of digital objects and construct them into a digital collection with ease. There is no limitation on what the digital objects can be, including video, audio, image in addition to text. The digital collections constructed by *Greenstone* can also be full-text searched and browsed, based on the metadata of each digital object within the collections.

A simple definition of metadata can be encapsulated as *“Data that describes other data”*. For example, we can assign title, creator, subject, description, and so forth to each digital object. There are many different metadata standards that have been developed. One of the most widely used is called "Dublin Core". Greenstone is agnostic towards metadata set, providing "Dublin Core" as a default.

In order to accommodate the diversity of digital objects and construct them into a digital collection, *Greenstone* utilises the concept of document parsing "plugins". Each plugin has different abilities for processing different digital objects.
For Example, a plugin *"PagedImagePlugin"* (Greenstone Plugins 2009) has been developed in the system to sustain the digital objects consisted with digitised document image and OCR text.

The plugin will associate each digitised image to a text file and give the user the option to view both sources of information. The drawback for building the newspaper collection in such a manner is that the processing of OCR images is assumed to, already have happened, this might not be of a great concern in the small size collection. However, it will be extremely time consuming and inefficient for the collection with greater scale. Even with the most cutting edge OCR technology we cannot guarantee the result will be one hundred percent accurate due to many different factors, and almost always the correction of the results is necessary by human editors in order to achieve the perfect result.

The functionality of the plugin can also be further extended based on the user's requirements. Once the collection has been built by Greenstone, it can then easily be made accessible through a variety of mediums, such as internet, CD-ROMs, DVDs and most recently even on an iPod (Bainbridge et al. 2008).

Chapter 2 - Background

There are several essential steps a digitised document must undergo for inclusion into a digital library: image pre-processing, analysing document layout. Each of steps will determine the quality of the end result in the digital library. Details of these steps are given in this chapter. In addition, a review of related previous projects is given. Following this we look at several possible framework candidates that can be used for building the framework.

2.1 Image-processing techniques

2.1.1 Binarisation

One of the fundamental techniques that is often utilised, before the process of OCR, is called binarisation. During the process of digitisation of the historic newspapers through either scanner or digital camera, common practice is to save the resultant image as greyscale. The reason for this is that the majority of newspapers (especially historic newspapers) are published in monochrome. It was not until recent decades that a significant amount of multicolour printing started to emerge in the contemporary newspapers.

The most straightforward algorithm for applying binarisation on the greyscale image, is to choose a threshold value such as 127 (the approximate midpoint from 255), and turn all the pixel values below the threshold to 1 (black), followed by changing the values above the threshold to 0 (white). The resulting image will then be binarised. This, algorithm however, is inadequate for many situations, as the threshold value might not be evenly distributed throughout the whole image. For example, it could be affected by uneven distribution of light, when the image is taken by the digital camera. In order to solve such deficiencies many other binarisation algorithms have been proposed — such as Sauvola (Sauvola, Pietikäinen 2000) and Otsu (Otsu 1979) — which present us with more complex methods that allow the threshold to be adaptive depending on different digitised images. This class of algorithm is frequently referred to as adaptive binarisation or adaptive thresholding. As a result, the algorithm provides more accurate binarisation to greyscale imaging.

Binarisation algorithm for coloured images can also be created in a similar manner. It can also be achieved by first converting the image into greyscale then processing the image with binarisation. Not only is binarisation one of the most commonly known image pre-processing techniques, but as well, takes an important role before OCR, as the majority of commercial OCR software requires the image to be binarised before it can recognise the characters within the image. Several other image processing algorithms also required the input image to be binarised beforehand.

2.1.2 De-skew

Skew often occurs during the process of image digitisation. For instance when we place a paper document on a scanner, most often the document will not be one hundred percent perpendicular to the edge of the scanner. Consequently digitised image will appear to have a certain amount skew to either the right or left. Similarly, if we capture an image with the digital camera, unless it is fixed with something like tripod the image will most probably result in skew.

Alternatively skew may already be present in a printed document. For instance, in the setting of the moveable type used when the historical newspapers were printed. Hence skew can easily result through such technology.

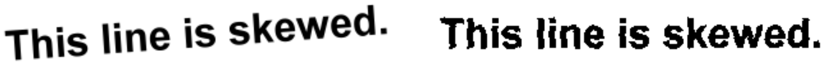

This line is skewed. This line is skewed.

Figure 2.1a: Example of line text de-skew (left: before, right: after).

In order to solve the problem of skew regardless of its cause, many algorithms have been developed for the purpose, such as the nearest neighbour clustering (Ávila, Lins 2005), Hough transform (A. Amin, S. Fisher 2000), Projection Profile (A. Bagdanov, J. Kanai 1997) and many more. Each de-skew algorithm has its own characteristics. Some perform better than others over different types of documents. The implementation of algorithms also has an effect on the performance (time spent, error rate) and results.

Figure 2.1a demonstrates a simple example of a de-skewing line of text. The image on left hand side illustrates a line of text with a few degrees of skew towards left — the image on the right hand side demonstrates the line of text after de-skew algorithm has been applied (based on the RAST algorithm described in Section 2.1.5).

Most de-skew algorithms are limited to dealing with one skew angle, (such as the image in Figure 2.1a) although the text is skewed each character is still positioned in a line. In such cases by correcting the line to horizontal or 180 degrees all the text will be de-skewed. However if each character is skewed in its own direction i.e. all the characters are not positioned in one line, or multiple skew exists in one document image, it will then require extra effort for correcting such skew.

Related studies has been done (Spitz 2003) which demonstrates a promising technique for de-skewing document images with multiple skew and even image distortions caused during digitisation.

De-skew has become one of the essential image processing techniques, especially for document images that contain text. It is also a crucial process before OCRing the image, as the OCR engine relies on comparing character image in order to distinguish between the characters. Hence applying OCR without de-skew will significantly decrease the accuracy of character recognition. In the majority of cases de-skew is necessary, prior to utilising further image processing techniques such as layout analysis algorithm or segmentation algorithm. Often implementation of algorithms will take the collocation of each character, sentence or paragraph into account. Therefore without de-skew, error rate could be increased to a great extent for many algorithms.

2.1.3 Noise-Removal

During the process of document image digitisation, noise could be created from the dust or spot attached to the digitisation device, such as the surface of the scanner or lens of the digital camera, et cetera. Typically types of noise called “salt and pepper” are often caused by lack of light and interruption of dark during the process of digitisation on optical devices.

Furthermore, spots can originate from the original document, in particular for the historic newspapers where significant amount of ink residue will be expected. For a period of time, microfilm has been mostly used for historic newspaper preservation. However through the period of time, the quality of the film, which maybe starting to degrade, will contribute to the noise during the digitisation.

It will be impractical and time consuming if one attempts to remove the noise from a collection of digitised document images. For that reason various noise removal algorithms have been developed. Algorithms such as non-linear based (Rudin, Osher, Fatemi 1992), or linear based (Song, Delp 1992), plus a variety of other algorithms have been developed or extended, based on these two categories. Both categories possess different characteristics and have often been used simultaneously or on top of each other.

A variety of noise removal functions can frequently be found in many commercial photos editing software such as Adobe Photoshop, and PhotoImpact, et cetera. They also provide in an open source (free distribution) photo editing software such as GIMP, Netpbm and ImageMagick, et cetera.

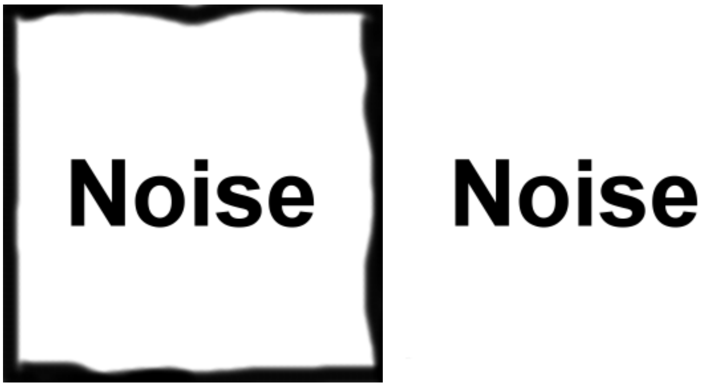


Figure 2.1b: Example of noise removal (left: before, right: after).

Figure 2.1b demonstrates a simple example for removing border noise from a digitised image using an algorithm developed throughout this thesis (more will be mentioned later on). Such border noises, often occurs during the process of image digitisation of the microfilm. Nevertheless, this algorithm presents only one of the solutions among many others, for noise removal.

In the processing of OCR, noise in the digitised image is an imperative factor. Any undesired noise within the range of the character will potentially increase the error rate of OCR. For example, if noise existed underneath the character "v", the OCR engine will very likely misinterpret it as character "y".
Layout analysis and segmentation algorithms too will be influenced by the noise, as some algorithms can be extremely sensitive to every component within the image, even the most insignificant one. Consequently, incorrect analysis and results might be given.

Noise removal is an essential image pre-processing technique, especially when dealing with an extensive amount of digitised images. It is tremendously time consuming if artificial processes are to be involved. Nevertheless, different algorithms can be further extended or developed to enable them to be used under different circumstances.

2.1.4 Document Layout Analysis

When we digitise a document, it becomes a digitised document image which can be displayed on the computer screen. Nonetheless, the image of a document is significantly different compared to a document in Microsoft Word or a document in a PDF file.

In the Word document we are able to define each page, paragraph, sentence, word and even characters with an incredible amount of detail and information. This information allows Word to distinguish different components within this document. For example, when we select a title and assign to it with heading properties in Word, its appearance will be changed according to the default settings in Word. Furthermore, when we insert an image in Word, we automatically tell the software it is an image. Such information is vital to Word, as it describes both the physical and logical structure of a document, components within the document and how they should be displayed. This information can also be used by Word for exporting the document as a different type of file in order for it to be utilised by other software.

Contrary to the above, such document structure information does not existed in the digitised document image. In general, image files only carry information such as, type of image file, dimension of the image (width and height), resolution (horizontal and vertical), and value of each pixel within the image, et cetera.

Therefore, techniques for extracting the information of document structures is often is referred to as Document Layout Analysis (DLA) or Document Image Analysis (DIA).

This process of automatically identifying and classifying the possible elements within a digitised document image and extracting them, is also referred to as document segmentation or automatic zoning.

Generally speaking, a digitised document can be constructed with many different types of elements. Types of the elements can vary depending on the structure of the document. Take the newspaper for example in the First international Newspaper Segmentation Contest (Gatos et al. 2001). Seven elements (Text, Title, Inverse Title (heading), Horizontal line, Vertical line, Photo and Graphic/Drawing) had been chosen for evaluating the structure of the newspapers.

These seven elements cover most of the components that we could find in the newspaper (both historic and contemporary). Elements can also be further refined depending on the circumstances, such as separated title into headings and sub headings, or text into columns, paragraphs, lines, words even characters. Nevertheless, not all elements are applicable to all newspaper structures. For instances photographs do not appear in the historic newspapers prior to 1897.

In 1897, *The New York Tribune* was the first newspaper that started to print halftone photographs, by adopting the photoengraving process that had been invented in 1860 in England and perfected by Federic E. Ives of Cornell University in 1886 (Kanungo, Allen 1999). Therefore we can say that prior to the invention of a photoengraving process, only hand sketch graphic and drawings would be expected in the newspapers.

In the past two decades, many algorithms have been developed in order to provide a possible solution for the requirement of DLA. Algorithms such as projection profile, connected components, smearing, Hough Transform, RAST, and XY-Cut, et cetera. Each technique has its strengths and weaknesses, depending on objectives and the implementation of the algorithm. Some perform better than others.

A review among several segmentation algorithms has been given (Cattoni, Coianiz et al. 1998). This review categorised the adopted algorithms based on different objectives, such as text segmentation, page segmentation et cetera. Each objective has its own purpose, and a brief description for the implementation of the algorithms for different research along with their advantages and disadvantages being mentioned.

Document segmentation is an important process to both OCR process and construction of a digitised document collection within DL. Imagine if we feed an image (without any line, photo, or drawing) containing one word, one sentence or one paragraph of text to the OCR software, most of the time the software should be able to translate the image into ASCII text without too much difficulty. However, if we feed the OCR software with a document image (with lines, photo, or drawing) consisting of certain layout structure such as historic newspapers, it would then be challenging for the OCR software to return ASCII text with high accuracy, as the lines or photos will probably cause recognition error, due to the fact that they are unable to be recognised by the software.

On the other hand, if we carry out the segmentation before passing on the image to OCR software, it will reduce the risks of inputting non-text images which might simply be ignored or output as gibberish by the OCR software. Moreover, we simplify the tasks for OCR software by providing one block of text image at a time, which guarantees certain recognition accuracy.

Furthermore, the layout information extracted after segmentation, can also be included when building the digital collection, such as formatting the ASCII text according to the original layout, or highlighting the keyword searched, as previously demonstrated in the *Papers Past* website. This will provide the user with additional supplementary support when accessing the digital collection.

Nevertheless, there is still room for the algorithm to be further enhanced, in order to provide better Document Layout Analysis for many other purposes.

In the following section, we will briefly introduce three segmentation algorithms (RAST, XY-Cut and Hough Transform) that have been utilised in our experiments, along with a brief of their history and how they work.

2.1.5 RAST Algorithm

RAST stands for *"Recognition by Adaptive Subdivision of Transformation Space"* (Breuel 1992). Compared with other algorithms RAST uses no heuristics nor looses potential solutions. It combines the idea of multi-resolution matching, Hough Transform, search-based recognition, and bounded error recognition. In (Breuel 1992) empirical data shows the RAST algorithm has a better performance than Hough Transform.

The basic idea of RAST, is to start with a transformation space, which contains all the transformation models that we need from the image. The algorithm then finds all the associations between the images and the model. The matching results are then evaluated from the associated sets. The RAST algorithm first starts with finding potential solution at large scale and refining and verifying them recursively into finer scale.

The RAST algorithm can be implemented for a diverse range of purposes, such as 3D object recognition from an image, facial recognition in modern digital camera, and document segmentation.

2.1.6 XY-Cut Algorithm

The initial idea of XY-Cut algorithm can be traced back to 1984 (Nagy, Seth 1984). The algorithm was presented in order to describe the basic layout for a digitised document image (by scanner). This algorithm was referred to as X-Y Tree by the authors. With the limited computer memory issue back then, the author proposed using tree as the data structure. Hence it was referred as X-Y Tree.
The primary initiative of this algorithm is to structure a digitised image into many rectangular blocks, with each node in the tree represented as a rectangular block obtained from its parent's node. The rectangular block can be obtained vertically (Y Cuts) or horizontally (X Cuts) depending on the pixel gaps (white space)

between each block in both X and Y directions.

Akin to RAST algorithm where the model needs to be defined, in XY-Cut algorithm, “rules” have to be set as well depending on the structure of the document, in order for the algorithm to determine where to perform the cutting.

As the result of this algorithm, the first node (root node) in the tree will represent the entire page image. Layout structure of the document will also be preserved according to the cutting sequences, with the exact location of each block and its size to be recorded. This process has often been implemented recursively throughout its development over the years, and has therefore been referred to as the Recursive XY-Cut algorithm.

Due to the fact that this algorithm is initially designed to recognise the layout structure of technical journals, and it has often been used in the manner of top-down, it is generally more suitable for documents with hierarchical structure such as technical journals, business records, and newspapers et cetera. Furthermore, the page is guaranteed to be fully segmented.

2.1.7 Hough Transform

This algorithm was invented by Richard O. Duda and Peter E. Hart in 1972 (Duda and Hart 1972) they referred to the algorithm as the generalised Hough Transform. Primarily this algorithm was invented for the needs of line and curve recognition from the digitised images. However, it has been made popular by Ballard (Ballard 1980) with his work of using Hough Transform for detecting different shapes within the image, such as circles, ellipses and triangles. Moreover, we can use Hough Transform as a universal transform, in order to detect any arbitrary shapes.

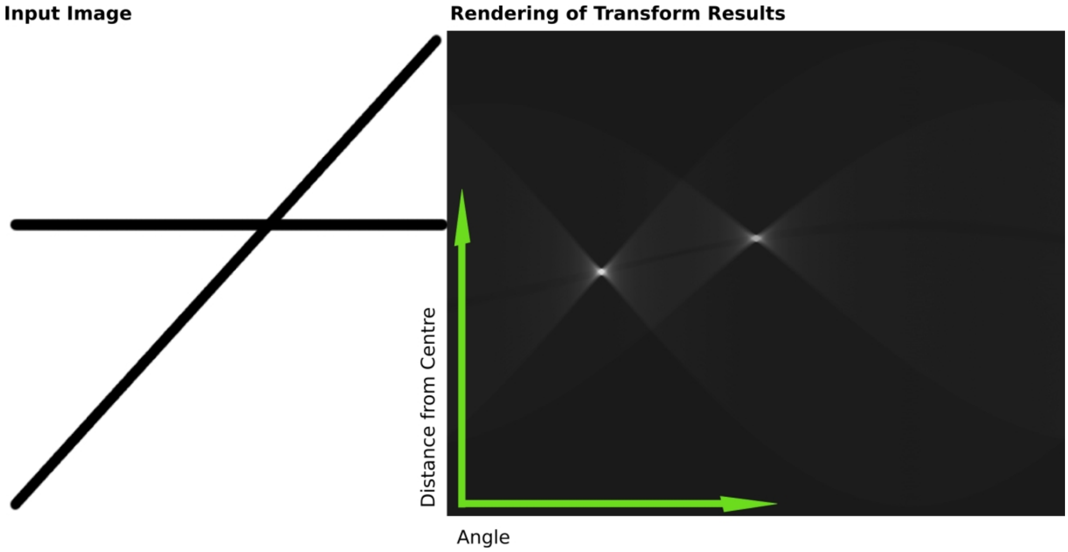


Figure 2.1c: Example of line transform into Hough space (GNU free documentation license).

The fundamental concept for line detection in Hough Transform is to convert the potential lines from the image into a two dimensional Hough space. Figure 2.1c demonstrates the transformation of two lines from an image into Hough space. Hough space is represented by two vectors (distances from the centre and an angles of the line), we can treat the Hough space as a two dimensional array of accumulators. Whenever the line passes through a pixel at the particular distances and angle, we increase the value of the correspondent accumulators. Consequently, the accumulators containing the highest value, indicates the potential lines on the image, which will also be represented as the brightest dots in the Hough space.

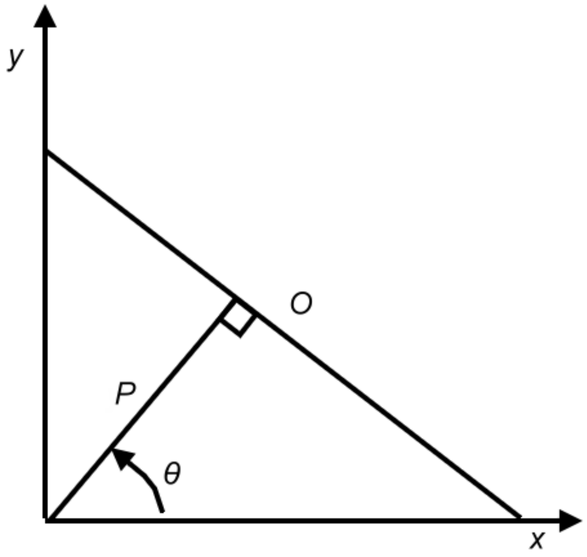


Figure 2.1d: Geometric representation of a line.

Hough Transform utilises an algebraic equation to validate the line. Figure 2.1d illustrates a geometric representation of line "O" within a two dimensional space. If we draw a perpendicular line "P" from the origin to line "O", by knowing "X" and "Y" we can then find out the angle of line "O" using the following equation.

$$P = x \cos \vartheta + y \sin \vartheta$$

Hough Transform can also be used for document skew detection, as it can accurately detect the angle of the lines appearing within an image to the smallest degree of precision (0.5 degree or even 0.1 degree if necessary). Alternatively, it is also a great algorithm for detecting the staves on the music sheets.

2.2 Related Document Image Analysis Projects

In the past decades there have been many projects focused on document image analysis and archiving. Each of them proposes different solutions for analysing and archiving their target documents and they tend to deal with one genre of document collection: newspaper, index card, journal, letters, and so on.

During the process of document analysis, user configurability of the settings used in the system, can have significant impact on the accuracy of the results. The prototype system has been built in such a manner in order to archive more than 500,000 index cards of Lepidoptera (Butterfilies and Moths) and Coleoptera (Beetles) for the UK Natural History Museum (Downton et al. 2006). The archive is then made searchable with validation and correction support through web-editing of the online digital archive. The system has the ability for image pre-processing, document analyse, OCR (Abbyy FineReader 6.0) and post-processing.

Five binarisation algorithms have been implemented into the system for image pre-processing. XY-Cut algorithm has been chosen for DIA as the index cards are of text block structure. Therefore the XY-Cut algorithm is more suitable for this type of document. OCR software is configured with specific dictionaries in order for more accurate word recognition. Regular expression is used during the post-processing for generating database strings for input to the online database.

The dimensions and structure of an index card are roughly equal. For example, “Species” has always been typed on the top left hand corner of the card. However, there could be slight differences between each card. Therefore a fuzzy configuration has been introduced into the system, which allows users to approximately define the region of the desired text block. The configuration can then be saved for the purpose of batch processes on other index cards with similar layouts.

Full-text access for historical newspapers has posed a potential problem from time to time, especially with the larger scale collections. A possible system can be build in order to provide universal access to historical newspapers with full-text searchable access (Kanungo and Allen 1999). It aims to extract the entire newspaper articles from zones of text, which back in 1999, to a great extent have not been considered by the research communities. The proposed system is involves three main functions - OCR module, Information Retrieval module and User Interface module. The potential problem of historic newspaper layouts have been mentioned in this research. Page layout has changed over time. In the early 1800s newspaper articles were typed very closely together with little space between lines and columns. The layout and fonts change from time to time and can be quite distinct according to different countries.

Kanungo and Allen built a small corpus from the paper copy of original papers of the Brooklyn Eagle for November 11, 1917, as well as from the microfilm which covers issues of the newspapers.

As much commercial OCR software fails to recognise text in highly degraded images, Kanungo and Allen decided to build a prototype OCR system using a commercial development kit. A number of Image pre-process techniques have also been included within the Information Retrieval module, such as de-skew, line removal, noise removal and finally with the XY-Cut algorithm for segmentation. After segmentation they classified each zone into text and non-text using a statistical decision tree. However it requires a dataset of images from manual segmentation and labelling.

Two type of interface have been suggested in the User Interface module. First interface is for the end user such as researchers and students. They can directly access the newspaper by a full-text search. Second an interface for corpus developers, which allow them to inspect and update OCR and segmentation errors.

In order for evaluating the performance of a system, the authors propose to create an abundant amount of corpora, which includes digitising the Negro Newspaper Collection as well as other ethnic newspapers around USA, in addition to 15 years of The New York Times. Ultimately, statistical stratified sampling methods, will be used to create a representative image dataset with ground-truth for the evaluation. Ground-truth can be defined as "an authentic sections on the document image contains desired information by human editor, in addition with its layout precisely corresponded to other sections both physically and logically."

Other than English, research of analysing and archiving historic newspapers has also been done in other languages, for example Greek (Gatos et al. 2000).

Authors presented the solution to problems related to newspaper image enhancement, segmentation into various items (title, text, image et cetera), and dataset from a large test bed of old newspaper issues.

During image pre-processing, an accurate and quick method has been used for image enhancement. Authors transform the binary image into greyscale then apply filtering using the slide window with a predefined threshold. This process is then repeated for both shrinking and swelling, in order to achieve satisfactory image enhancement. In terms of de-skew, fast Hough Transform has been chosen for its enhanced performance over the normal Hough Transform. A technique has been developed by authors in combination with fast Hough Transform for efficient skew detection, which they referred to as Block Hough Transform.

In terms of page segmentation, a new technique has been proposed, based on gradual extraction of the components in the order of lines, images and drawings, background lines, special symbols, text and title block. Identification of potential images or drawings have been performed using the Run Length Smoothing Algorithm. The segment is then extracted using Connected Component Analysis. Finally the segment is analysed by Fast Fourier Transform to confirm the existence of images or drawings. Special emphasis on article identification and reconstruction, has also been mentioned by the authors, although these issues have been addressed by many researchers. However majority of them tend to focus on scientific journals, which makes the results less important for the layout of newspapers, as they have different layout. The authors approach towards this issue is to examine the relationships between each segment of the newspaper layout.

The authors integrated an OCR module developed from earlier work in combination with overlapping Gaussian masks based algorithm for character segmentation, and the results of 99.7% recognition rate have been achieved.

The experiment used the newspaper “TO VIMA” which has been published by Lambrakis Press S.A. since 1922 to the present time, and who owns a large collection of newspapers consisting of 1.3 million pages, dating from the 1890 up to the present. Experimental results are obtained from a test set of 100 pages from newspaper “TO VIMA” published between 1965 to 1974.

There are many other similar projects related to document analysis that have been done other than newspapers. However, with different document layouts the analysis approach might need to be different, despite many documents sharing various commonalities. Much research still needs to be done in order to distinguish their differences and enhance the results of the segmentation.

2.3 Implementation Framework

There are many possible frameworks that we can utilise for implementing a system for the purpose of image pre-processing, document analysis and image OCR.

One of the potential open source frameworks *Gamera* (Choudhury et al. 2006) has the ability for building systems that can extract information from digitised two-dimensional documents. It can be used to support many different types of documents. *Gamera* particularly focuses on recognition of the documents containing cultural heritage materials, such as sheet music, medieval manuscripts, et cetera. Such documents can be ancient and fail to be recognised by the majority of commercial OCR software. *Gamera* combines with a programming library using GUI tools which allows users to construct a system for recognition of a particular document. *Gamera* continues to develop since 2001 and uses the approach to providing an open, learning based, flexible, and distribution system.

Alternatively a possible framework can be developed by using *ImageJ*. It is an open source Java based image processing software developed by Wayne Rasband at the Nation Institute of Health. It can be run on any platform with a Java 1.4 or later virtual machine. It has the ability to display, edit, analyse and process images with various formats. In addition it provides various image enhancement functions, such as binarisation, edge detection, and median filtering et cetera.
It is initially designed on open architecture and is capable of providing extensibility through Java plugins. Therefore image analysing and processing plugins can be developed which promise the user being able to solve the majority of image analysing and processing problems. It is currently known as the fastest pure Java image processing software, with the ability to filter 40 million pixels per second.

Most recently an open source OCR system called *Ocropus* has been released (Breuel, Kaiserlautern 2007). First release (version 0.1.1) of *Ocropus* was in October 2007 and they have continued to release newer versions of the system. At the time of writing this thesis, version 0.3 has been released. The system is being developed with support from Google and with the primary developers from IUPR (Image Understanding and Pattern Recognition) Research Group. The system is aimed to provide a flexible platform for both the research community and commercial researchers. Differing from most commercial OCR software, *Ocropus* not only includes an open source OCR engine *Tesseract* (mature OCR engine developed by Hewlett and Packard), but also provides users with numerous interfaces, such as for document analysis, image pre-processing (binarisation, noise removal, de-skew), pages segmentations and language modelling et cetera. *Ocropus* is developed using C++ (programming language). However Lua scripting language has been integrated for it to be used efficiently used by the domain experts and other users.

There is currently three page segmentation algorithms implemented in *Ocropus*, including RAST, XY-Cut and Voronoi-based. However, more algorithms might be implemented in due course. Extensibility is also one of the main features of *Ocropus*, as new algorithms can be easily incorporated into the system. In addition *Ocropus* can be used to analyse documents in languages other than the English, as we can train *Tesseract* to recognised characters other than the English alphabet.

We have chose *Ocropus* as the implementation framework for our experiments, due to its extensibility and built-in OCR engine. In addition along with certain algorithms that have already been introduced into this system it makes *Ocropus* a potential candidate to compare to others.

Chapter 3 - Software Implementation and Corpus

In this chapter we describe some of the working environment for *Ocropus*. Implemented algorithms are also be mentioned, in addition to the historic newspaper corpus used.

Ocropus is primarily developed on *Ubuntu* (an open source Linux-based operating system). Nevertheless, it can be installed with certain external software on several other platforms such as Mac OS X, Solaris, MS Windows and CentOS. However, as it was developed under *Ubuntu* it therefore cannot be guaranteed to work under other platforms.

For our implementation, we have chosen to work with *Ocropus version 0.2* under the Microsoft Windows XP environment. In order for it to be working in Windows, we installed *Cygwin* (open source software that creates a Linux like environment for Windows) version 1.5. *Cygwin* allows us to compile the *Ocropus* source code in order for it to work under Windows.

3.1 Trim Edge Algorithm

Ocropus is currently implemented with a document cleanup algorithm for removing the noises on the image. The algorithm, using black filter followed by a connected component analysis resulting in a clean output by white filter.

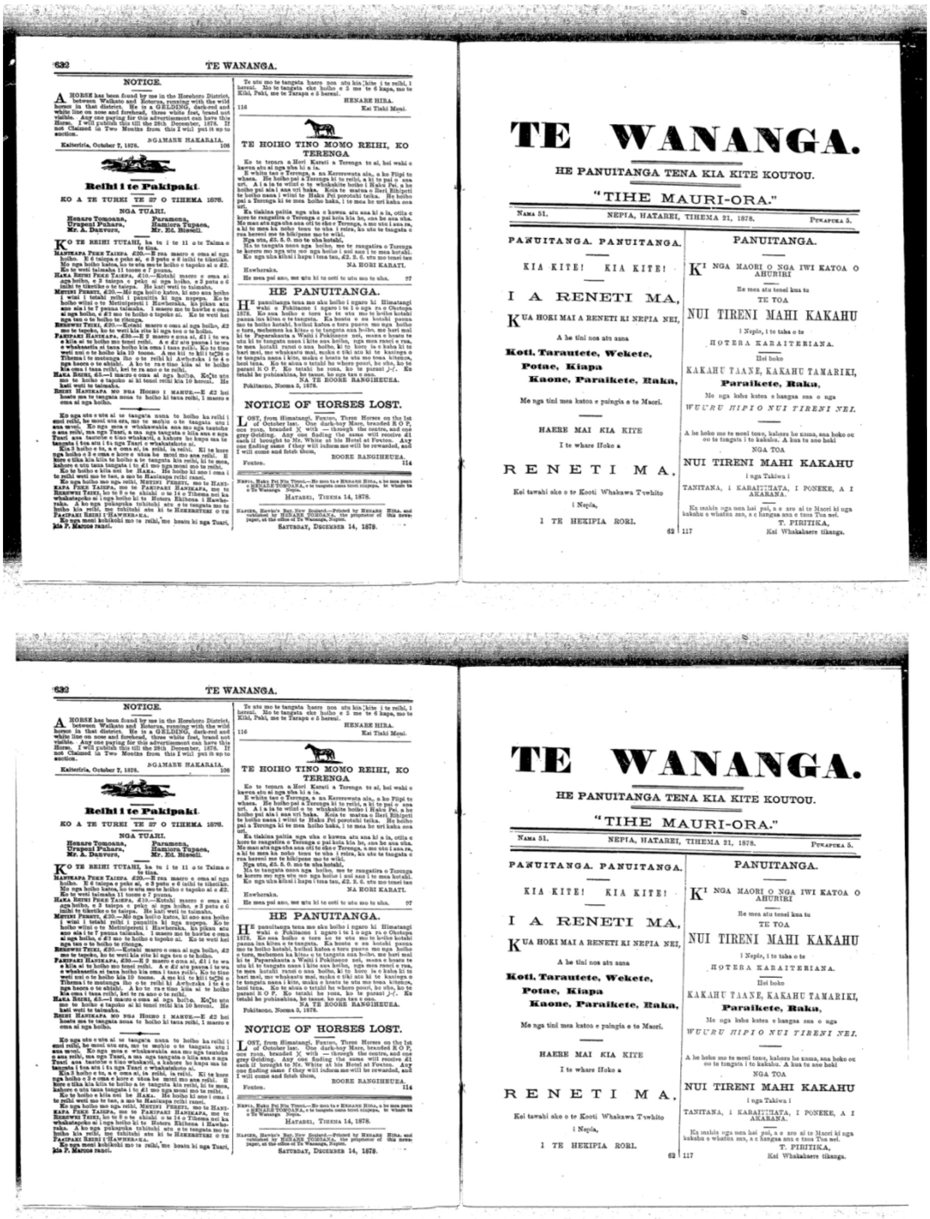
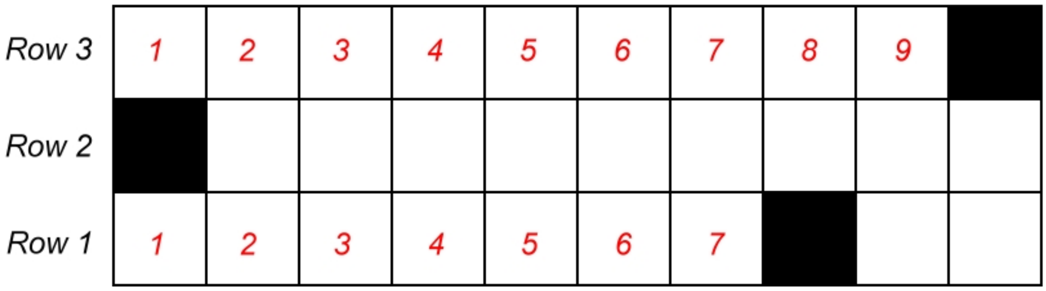


Figure 3.1a: Image with applied document cleanup algorithm (top: before, bottom: after).

Figure 3.1a demonstrates the results after applying a document cleanup algorithm, implemented in *Ocropus* to an image with noises around the borders which were produced during the digitisation process. As we can see the algorithm removes the majority of noises from left and right hand side of the image. However, great amounts of noise remain at the top of the image. The result from the algorithm is not satisfactory enough to meet our standards, which is to remove the majority of the border noises.

These kinds of noise can be commonly found from the images digitised from microfilm. Therefore in order to meet our requirements in addition to creating an automated process for removing border noises, we implemented a trim edge algorithm in using the scripting language embedded in *Ocropus*, Lua.

In order to remove the border, the algorithm first examines the most promising row position for each direction in the order of left, top, right and bottom. Once the row position has been identified we then remove the noise by changing all the pixel values before that row to 255 (white). This algorithm takes the assumption that the images have white background and black font, which most documents do.



Row 3	1	2	3	4	5	6	7	8	9	■
Row 2	■									
Row 1	1	2	3	4	5	6	7	■		

Figure 3.1b: Counting of maximum continued white pixels.

The most promising row position is determined by examining the maximum value of continued white pixels in the current row and the row after next. As demonstrated in Figure 3.1b, assumes we want to find the most promising row starting from bottom of the image (going upwards). We then find out the maximum value of continued white pixels for “Row 1” and “Row 2”. If we find a black pixel (possibly noise) we then reset the counter to zero followed by counting the next continued white pixel. When we reach the last pixel we then return a result showing the maximum value found.

Both maximum values are then divided by the total number of pixels, in this case width of the image. If the divided result equals one, this indicates both rows contain one hundred percent white pixels, which is the optimal result possible. In such a case we can then take the outermost row as the most promising. Nevertheless, in real practice, most of the time we might not be able to find the rows containing one hundred percent white pixels, due to the possible noises. If this is the case, we then look for the second best ratio of 90%, 80%, and 70% respectively. For instance in Figure 3.1b the best option will be “Row 3”, although it did not achieve hundred percent, but along with “Row 1” they both reach the ratio of 70%.

The main reason for omitting the row between one and three, is for us to reduce the calculations needed for the algorithm, and improve the performance. We can further omit more rows depending on the solutions. Nevertheless, we need to examine two rows simultaneously in order to find enough evidence for the most promising row.

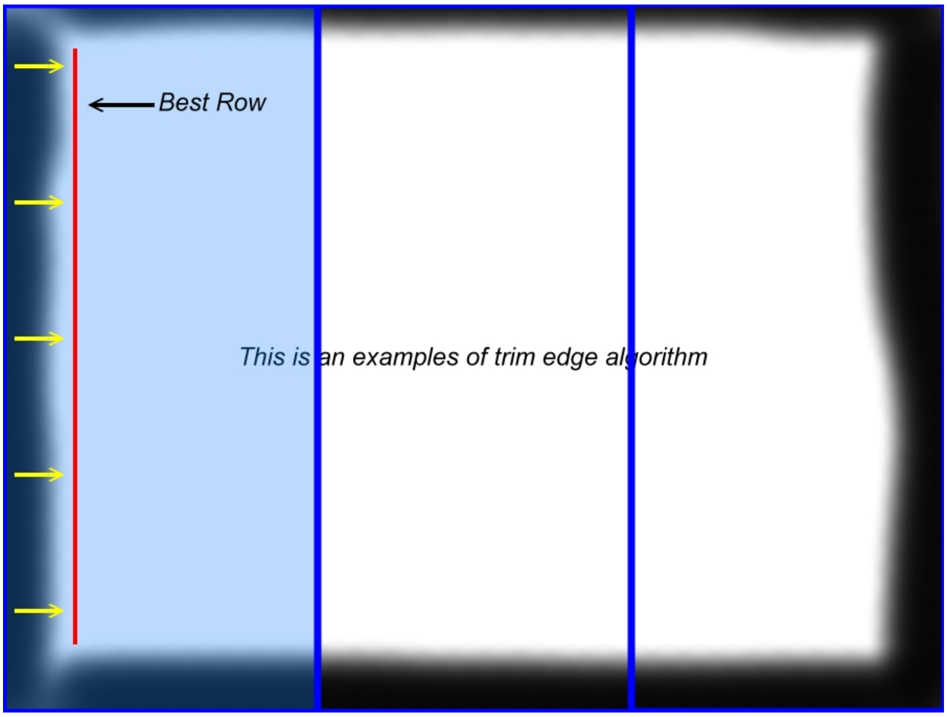
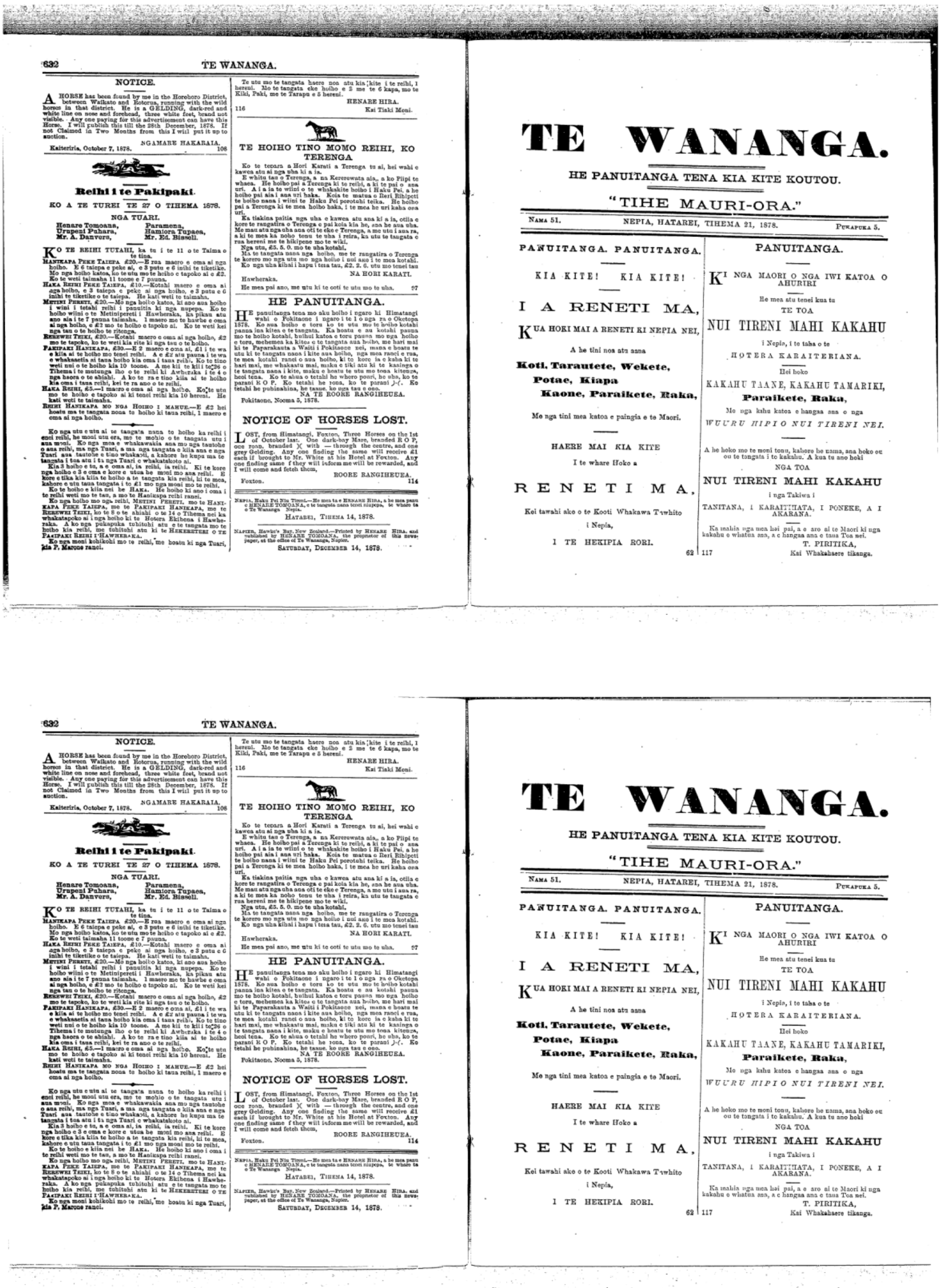


Figure 3.1c: Search for the best row from left hand side.

In order to further enhance the performance of the algorithm we take each direction into consideration. As shown in Figure 3.1c, when we want to trim off the border noises on left hand side, we first divided the image into three sections. By doing this we reduce the total search area by 66%. This also prevents the algorithm from searching through the entire image if it fails to find the most promising row in the first section. Secondly, we examine the value of white continued pixels according to the directions from left to right. Once the best row is found we then erase the noises before the row. Equivalent procedure is then performed for other edges of the image respectively. While it is possible to divide the images into smaller sections for a more enhanced performance, dividing by three appears to produce with better results.



632 TE WANANGA.

NOTICE.

A HORSE has been found by me in the Horohoro District, between Waikato and Rotorua, running with the wild horses in that district. He is a GELDING, dark-red and white line on nose and forehead, three white feet, brand not visible. Any one paying for this advertisement can have this Horse. I will publish this till the 28th December, 1878. If not Claimed in Two Months from this I will put it up to auction.

NGAMARE HAKARAIA.

Kaiteriria, October 7, 1878. 108

Reihi i te Pakipaki.

KO A TE TUREI TE 27 O TIHEMA 1878.

NGA TUARI.

Henare Tomoana, Urupeni Puhara, Mr. A. Danvers, Paramena, Hamiora Tupaea, Mr. Ed. Bissell.

KO TE REIHI TUTAHI, ka tu i te 11 o te Taima o te tina.

HANIKAPA PEKE TAIEPA £20.—E rua maero e oma ai nga hoiho. E 6 taiepa e peke ai, e 3 putu e 6 inihi te tiketike. Mo nga hoiho katoa, ko te utu mo te hoiho e tapoko ai e £2. Ko te weti taimaha 11 toone e 7 pauna.

HAKA REIHI PEKE TAIEPA, £10.—Kotahi maero e oma ai nga hoiho, e 3 taiepa e peke ai nga hoiho, e 3 putu e 6 inihi te tiketike o te taiepa. He kati weti te taimaha.

METINI PERETI, £20.—Mo nga hoiho katoa, ki ano ana hoiho i wini i tetahi reihi i panuitia ki nga nupepa. Ko te hoiho wiini o te Metinipereti i Hawheraka, ka pikau atu ano aia i te 7 pauna taimaha. 1 maero me te hawhe e oma ai nga hoiho, e £2 mo te hoiho e tapoko ai. Ko te weti kei nga tau o te hoiho te ritenga.

REREWEI TEIKI, £20.—Kotahi maero e oma ai nga hoiho, £2 mo te tapoko, ko te weti kia rite ki nga tau o te hoiho.

PAKIPAKI HANIKAPA, £30.—E 2 maero e oma ai, £1 i te wa e kiia ai te hoiho mo tenei reihi. A e £2 atu pauna i te wa e whakaaetia ai taua hoiho kia oma i taua reihi. Ko te tino weti nui o te hoiho kia 10 toone. A me kii te kii i te 26 o Tihema i te mutunga iho o te reihi ki Awheraka i te 4 o nga haora o te ahiahi. A ko te ra e tino kiia ai te hoiho kia oma i taua reihi, kei te ra ano o te reihi.

HAKA REIHI, £5.—1 maero e oma ai nga hoiho. Ko te utu mo te hoiho e tapoko ai ki tenei reihi kia 10 hereni. He kati weti te taimaha.

REIHI HANIKAPA MO NGA HOIHO I MAHUE.—E £2 hei hoatu ma te tangata nona te hoiho ki taua reihi, 1 maero e oma ai nga hoiho.

Ko nga utu e utu ai te tangata nana te hoiho ka reihi i enei reihi, he moni utu era, me te mohio o te tangata utu i ana moni. Ko nga mea e whakawakia ana mo nga tautohe o ana reihi, ma nga Tuari, a ma nga tangata e kiia ana e nga Tuari ana tautohe e tino whakapti, a kahore he kupu ma te tangata i tua atu i ta nga Tuari e whakatakoto ai.

Kia 3 hoiho e tu, a e oma ai, ia reihi, ia reihi. Ki te kore nga hoiho e 3 e oma e kore e utua he moni mo ana reihi. E kore e tika kia kiia te hoiho a te tangata kia reihi, ki te mea, kahore e utu taua tangata i te £1 mo nga moni mo te reihi.

Ko te hoiho e kiia nei he HAKA. He hoiho ki ano i oma i te reihi weti mo te tau, a mo te Hanikapa reihi ranei.

Ko nga hoiho mo nga reihi, METINI PERETI, mo te HANIKAPA PEKE TAIEPA, me te PAKIPAKI HANIKAPA, me te REREWEI TEIKI, ko te 8 o te ahiahi o te 14 o Tihema nei ka whakatapoko ai i nga hoiho ki te Hotera Ekihena i Hawheraka. A ko nga pukapuka tuhituhi atu e te tangata mo te hoiho kia reihi, me tuhituhi atu ki te HEKERETERI O TE PAKIPAKI REIHI I HAWHERAKA.

Ko nga moni kohikohi mo te reihi, me hoatu ki nga Tuari, kia P. Marone ranei.

Te utu mo te tangata haere noa atu kia kite i te reihi, 1 hereni. Mo te tangata eke hoiho e 2 me te 6 kapa, mo te Kiki, Paki, me te Tarapu e 5 hereni.

HENARE HIRA.

116 Kai Tiaki Moni.

TE HOIHO TINO MOMO REIHI, KO TERENGA

Ko te tepara a Hori Karati a Terenga tu ai, hei wahi e kawea atu ai nga uha ki a ia.

E whitu tau o Terenga, a na Kererewata aia, a ko Piipi te whaea. He hoiho pai a Terenga ki te reihi, a ki te pai o ana uri. A i a ia te wiini o te whakakite hoiho i Haku Pei, a he hoiho pai aia i ana uri haka. Koia te matua o Reri Ribipoti te hoiho nana i wiini te Haku Pei porotuhi teika. He hoiho pai a Terenga ki te mea hoiho haka, i te mea he uri kaha ona uri.

Ka tiakina paitia nga uha e kawea atu ana ki a ia, otiia e kore te rangatira o Terenga e pai koia kia he, ana he aua uha. Me mau atu nga uha ana oti te eke e Terenga, a me utu i ana ra, a ki te mea ka noho tonu te uha i reira, ka utu te tangata e rua hereni me te hikipene mo te wiki.

Nga utu, £5. 5. 0. mo te uha kotahi.

Ma te tangata nana nga hoiho, me te rangatira o Terenga te korero mo nga utu mo nga hoiho i nui ake i te mea kotahi.

Ko nga uha kihai i hapu i tera tau, £2. 2. 6. utu mo tenei tau

NA HORI KARATI.

Hawheraka.

He mea pai ano, me utu ki te coti te utu mo te uha. 97

HE PANUITANGA.

HE panuitanga tena mo aku hoiho i ngaro ki Himatangi wahi o Pokitaone i ngaro i te 1 o nga ra o Oketopa 1878. Ko aua hoiho e toru ko te utu mo te hoiho kotahi pauna ina kitea e te tangata. Ka hoatu e au kotahi pauna mo te hoiho kotahi, huihui katoa e toru pauna mo nga hoiho e toru, mehemea ka kitea e te tangata aua hoiho, me hari mai ki te Paparakauta a Waiti i Pokitaone nei, mana e hoatu te utu ki te tangata nana i kite aua hoiho, nga mea ranei e rua, te mea kotahi ranei o aua hoiho, ki te kore ia e kaha ki te hari mai, me whakaatu mai, maku e tiki atu ki te kainga o te tangata nana i kite, maku e hoatu te utu mo tona kitenga, heoi tena. Ko te ahua o tetahi he whero pouri, he uha, ko te parani R O P. Ko tetahi he rona, ko te parani)-(. Ko tetahi he puhinahina, he taane, ko nga tau e ono.

NA TE ROORE RANGIHEUEA.

Pokitaone, Noema 5, 1878.

NOTICE OF HORSES LOST.

LOST, from Himatangi, Foxton, Three Horses on the 1st of October last. One dark-bay Mare, branded R O P, one roan, branded)(with — through the centre, and one grey Gelding. Any one finding the same will receive £1 each if brought to Mr. White at his Hotel at Foxton. Any one finding same if they will inform me will be rewarded, and I will come and fetch them,

ROORE RANGIHEUEA.

Foxton. 114

NEPIA, Haku Pei Niu Tireni.—He mea ta e HENARE HIRA, a he mea panui e HENARE TOMOANA, e te tangata nana tenei niupepa, te whare ta o Te Wananga Nepia.

HATAREI, TIHEMA 14, 1878.

NAPIER, Hawke's Bay, New Zealand.—Printed by HENARE HIRA, and published by HENARE TOMOANA, the proprietor of this newspaper, at the office of Te Wananga, Napier.

SATURDAY, DECEMBER 14, 1878.

TE WANANGA.

HE PANUITANGA TENA KIA KITE KOUTOU.

"TIHE MAURI-ORA."

NAMA 51. NEPIA, HATAREI, TIHEMA 21, 1878. PUKAPUKA 5.

PANUITANGA. PANUITANGA.

KIA KITE! KIA KITE!

I A RENETI MA,

KUA HOKI MAI A RENETI KI NEPIA NEI,

A he tini noa atu aana

Koti, Tarautete, Wekete,
Potae, Kiapa
Kaone, Paraikete, Raka,

Me nga tini mea katoa e paingia e te Maori.

HAERE MAI KIA KITE

I te whare Hoko a

RENETI MA,

Kei tawahi ake o te Kooti Whakawa Tawhito

i Nepia,

I TE HEKIPIA RORI.

62

PANUITANGA.

KI NGA MAORI O NGA IWI KATOA O AHURIRI

He mea atu tenei kua tu

TE TOA

NUI TIRENI MAHI KAKAHU

i Nepia, i te taha o te

HOTERA KARAITERIANA.

Hei hoko

KAKAHU TAANE, KAKAHU TAMARIKI,

Paraikete, Raka,

Me nga kahu katoa e hangaa ana o nga

WUURU HIPI O NUI TIRENI NEI.

A he hoko mo te moni tonu, kahore he nama, ana hoko ou ou te tangata i te kakahu. A kua tu ano hoki

NGA TOA

NUI TIRENI MAHI KAKAHU

i nga Takiwa i

TANITANA, I KARAITITIATA, I PONEKE, A I AKARANA.

Ka mahia nga mea hei pai, a e aro ai te Maori ki nga kakahu e whatua ana, a e hangaa ana e taua Toa nei.

T. PIRITIKA,

117 Kai Whakahaere tikanga.

632 TE WANANGA.

NOTICE.

A HORSE has been found by me in the Horohoro District, between Waikato and Rotorua, running with the wild horses in that district. He is a GELDING, dark-red and white line on nose and forehead, three white feet, brand not visible. Any one paying for this advertisement can have this Horse. I will publish this till the 28th December, 1878. If not Claimed in Two Months from this I will put it up to auction.

NGAMARE HAKARAIA.

Kaiteriria, October 7, 1878. 108

Reihi i te Pakipaki.

KO A TE TUREI TE 27 O TIHEMA 1878.

NGA TUARI.

Henare Tomoana, Urupeni Puhara, Mr. A. Danvers, Paramena, Hamiora Tupaea, Mr. Ed. Bissell.

KO TE REIHI TUTAHI, ka tu i te 11 o te Taima o te tina.

HANIKAPA PEKE TAIEPA £20.—E rua maero e oma ai nga hoiho. E 6 taiepa e peke ai, e 3 putu e 6 inihi te tiketike. Mo nga hoiho katoa, ko te utu mo te hoiho e tapoko ai e £2. Ko te weti taimaha 11 toone e 7 pauna.

HAKA REIHI PEKE TAIEPA, £10.—Kotahi maero e oma ai nga hoiho, e 3 taiepa e peke ai nga hoiho, e 3 putu e 6 inihi te tiketike o te taiepa. He kati weti te taimaha.

METINI PERETI, £20.—Mo nga hoiho katoa, ki ano ana hoiho i wini i tetahi reihi i panuitia ki nga nupepa. Ko te hoiho wiini o te Metinipereti i Hawheraka, ka pikau atu ano aia i te 7 pauna taimaha. 1 maero me te hawhe e oma ai nga hoiho, e £2 mo te hoiho e tapoko ai. Ko te weti kei nga tau o te hoiho te ritenga.

REREWEI TEIKI, £20.—Kotahi maero e oma ai nga hoiho, £2 mo te tapoko, ko te weti kia rite ki nga tau o te hoiho.

PAKIPAKI HANIKAPA, £30.—E 2 maero e oma ai, £1 i te wa e kiia ai te hoiho mo tenei reihi. A e £2 atu pauna i te wa e whakaaetia ai taua hoiho kia oma i taua reihi. Ko te tino weti nui o te hoiho kia 10 toone. A me kii te kii i te 26 o Tihema i te mutunga iho o te reihi ki Awheraka i te 4 o nga haora o te ahiahi. A ko te ra e tino kiia ai te hoiho kia oma i taua reihi, kei te ra ano o te reihi.

HAKA REIHI, £5.—1 maero e oma ai nga hoiho. Ko te utu mo te hoiho e tapoko ai ki tenei reihi kia 10 hereni. He kati weti te taimaha.

REIHI HANIKAPA MO NGA HOIHO I MAHUE.—E £2 hei hoatu ma te tangata nona te hoiho ki taua reihi, 1 maero e oma ai nga hoiho.

Ko nga utu e utu ai te tangata nana te hoiho ka reihi i enei reihi, he moni utu era, me te mohio o te tangata utu i ana moni. Ko nga mea e whakawakia ana mo nga tautohe o ana reihi, ma nga Tuari, a ma nga tangata e kiia ana e nga Tuari ana tautohe e tino whakapti, a kahore he kupu ma te tangata i tua atu i ta nga Tuari e whakatakoto ai.

Kia 3 hoiho e tu, a e oma ai, ia reihi, ia reihi. Ki te kore nga hoiho e 3 e oma e kore e utua he moni mo ana reihi. E kore e tika kia kiia te hoiho a te tangata kia reihi, ki te mea, kahore e utu taua tangata i te £1 mo nga moni mo te reihi.

Ko te hoiho e kiia nei he HAKA. He hoiho ki ano i oma i te reihi weti mo te tau, a mo te Hanikapa reihi ranei.

Ko nga hoiho mo nga reihi, METINI PERETI, mo te HANIKAPA PEKE TAIEPA, me te PAKIPAKI HANIKAPA, me te REREWEI TEIKI, ko te 8 o te ahiahi o te 14 o Tihema nei ka whakatapoko ai i nga hoiho ki te Hotera Ekihena i Hawheraka. A ko nga pukapuka tuhituhi atu e te tangata mo te hoiho kia reihi, me tuhituhi atu ki te HEKERETERI O TE PAKIPAKI REIHI I HAWHERAKA.

Ko nga moni kohikohi mo te reihi, me hoatu ki nga Tuari, kia P. Marone ranei.

Te utu mo te tangata haere noa atu kia kite i te reihi, 1 hereni. Mo te tangata eke hoiho e 2 me te 6 kapa, mo te Kiki, Paki, me te Tarapu e 5 hereni.

HENARE HIRA.

116 Kai Tiaki Moni.

TE HOIHO TINO MOMO REIHI, KO TERENGA

Ko te tepara a Hori Karati a Terenga tu ai, hei wahi e kawea atu ai nga uha ki a ia.

E whitu tau o Terenga, a na Kererewata aia, a ko Piipi te whaea. He hoiho pai a Terenga ki te reihi, a ki te pai o ana uri. A i a ia te wiini o te whakakite hoiho i Haku Pei, a he hoiho pai aia i ana uri haka. Koia te matua o Reri Ribipoti te hoiho nana i wiini te Haku Pei porotuhi teika. He hoiho pai a Terenga ki te mea hoiho haka, i te mea he uri kaha ona uri.

Ka tiakina paitia nga uha e kawea atu ana ki a ia, otiia e kore te rangatira o Terenga e pai koia kia he, ana he aua uha. Me mau atu nga uha ana oti te eke e Terenga, a me utu i ana ra, a ki te mea ka noho tonu te uha i reira, ka utu te tangata e rua hereni me te hikipene mo te wiki.

Nga utu, £5. 5. 0. mo te uha kotahi.

Ma te tangata nana nga hoiho, me te rangatira o Terenga te korero mo nga utu mo nga hoiho i nui ake i te mea kotahi.

Ko nga uha kihai i hapu i tera tau, £2. 2. 6. utu mo tenei tau

NA HORI KARATI.

Hawheraka.

He mea pai ano, me utu ki te coti te utu mo te uha. 97

HE PANUITANGA.

HE panuitanga tena mo aku hoiho i ngaro ki Himatangi wahi o Pokitaone i ngaro i te 1 o nga ra o Oketopa 1878. Ko aua hoiho e toru ko te utu mo te hoiho kotahi pauna ina kitea e te tangata. Ka hoatu e au kotahi pauna mo te hoiho kotahi, huihui katoa e toru pauna mo nga hoiho e toru, mehemea ka kitea e te tangata aua hoiho, me hari mai ki te Paparakauta a Waiti i Pokitaone nei, mana e hoatu te utu ki te tangata nana i kite aua hoiho, nga mea ranei e rua, te mea kotahi ranei o aua hoiho, ki te kore ia e kaha ki te hari mai, me whakaatu mai, maku e tiki atu ki te kainga o te tangata nana i kite, maku e hoatu te utu mo tona kitenga, heoi tena. Ko te ahua o tetahi he whero pouri, he uha, ko te parani R O P. Ko tetahi he rona, ko te parani)-(. Ko tetahi he puhinahina, he taane, ko nga tau e ono.

NA TE ROORE RANGIHEUEA.

Pokitaone, Noema 5, 1878.

NOTICE OF HORSES LOST.

LOST, from Himatangi, Foxton, Three Horses on the 1st of October last. One dark-bay Mare, branded R O P, one roan, branded)(with — through the centre, and one grey Gelding. Any one finding the same will receive £1 each if brought to Mr. White at his Hotel at Foxton. Any one finding same if they will inform me will be rewarded, and I will come and fetch them,

ROORE RANGIHEUEA.

Foxton. 114

NEPIA, Haku Pei Niu Tireni.—He mea ta e HENARE HIRA, a he mea panui e HENARE TOMOANA, e te tangata nana tenei niupepa, te whare ta o Te Wananga Nepia.

HATAREI, TIHEMA 14, 1878.

NAPIER, Hawke's Bay, New Zealand.—Printed by HENARE HIRA, and published by HENARE TOMOANA, the proprietor of this newspaper, at the office of Te Wananga, Napier.

SATURDAY, DECEMBER 14, 1878.

TE WANANGA.

HE PANUITANGA TENA KIA KITE KOUTOU.

"TIHE MAURI-ORA."

NAMA 51. NEPIA, HATAREI, TIHEMA 21, 1878. PUKAPUKA 5.

PANUITANGA. PANUITANGA.

KIA KITE! KIA KITE!

I A RENETI MA,

KUA HOKI MAI A RENETI KI NEPIA NEI,

A he tini noa atu aana

Koti, Tarautete, Wekete,
Potae, Kiapa
Kaone, Paraikete, Raka,

Me nga tini mea katoa e paingia e te Maori.

HAERE MAI KIA KITE

I te whare Hoko a

RENETI MA,

Kei tawahi ake o te Kooti Whakawa Tawhito

i Nepia,

I TE HEKIPIA RORI.

62

PANUITANGA.

KI NGA MAORI O NGA IWI KATOA O AHURIRI

He mea atu tenei kua tu

TE TOA

NUI TIRENI MAHI KAKAHU

i Nepia, i te taha o te

HOTERA KARAITERIANA.

Hei hoko

KAKAHU TAANE, KAKAHU TAMARIKI,

Paraikete, Raka,

Me nga kahu katoa e hangaa ana o nga

WUURU HIPI O NUI TIRENI NEI.

A he hoko mo te moni tonu, kahore he nama, ana hoko ou ou te tangata i te kakahu. A kua tu ano hoki

NGA TOA

NUI TIRENI MAHI KAKAHU

i nga Takiwa i

TANITANA, I KARAITITIATA, I PONEKE, A I AKARANA.

Ka mahia nga mea hei pai, a e aro ai te Maori ki nga kakahu e whatua ana, a e hangaa ana e taua Toa nei.

T. PIRITIKA,

117 Kai Whakahaere tikanga.

Figure 3.1d: Result comparisons between document clean and trim edge algorithms (top: document clean, bottom: trim edge).

Figure 3.1d illustrates the resulting comparison between two algorithms applied to the same document images (Original image in Figure 3.1c). As it demonstrates trim edge algorithm removal of extra thin strips of noise on the left and right hand side compared to the document clean algorithm. In addition, majority of salt and pepper like noises have also been removed from the top. Nevertheless, this algorithm suggests only one possible way for removing border noises. There may be many other possibilities that could be implemented to achieve similar results.

3.2 Hough Transform Algorithm

As mentioned previously in Chapter 2, Hough Transform is well known for its ability of line detection in an image, even with lines that are broken or have a gap in between, due to various reasons such as a printing mistake, deterioration of the newspaper, et cetera.

Throughout this thesis our focus is based on the historical newspapers. It appears that the layout of historical newspapers (and even contemporary newspapers) often contain certain amounts of lines (both horizontal and vertical). We believe that taking these lines into consideration during the segmentation process, will greatly improve the results for the newspaper-like type of documents.

For this reason, we have implemented a Hough Transform based algorithm for historical newspaper segmentation. Furthermore with a certain amount of alteration, based on the algorithm, we implemented an additional algorithm for the purpose of removing the line detected from the original image. Prior to Hough Transform, an image was firstly been binarised. We then calculated the maximum r (rMax) distances for that particular image. rMax represents the maximum distances allowed for any line from one original point on the image.

It is equivalent to the vector of “distances from centre” in Figure 2.1c.

To find rMax we use Pythagoras’ Theorem of:

$$c = \sqrt{a^2 + b^2}$$

Where “c” equals to rMax, “a” equals to halve of the image width and “b” equals to halve of the image height. By halving both of the dimensions, we made the rMax distance to the centre of the image, which will allow for a more straightforward calculation later on.

A two dimensional array is then created to be used as an accumulator, with rMax in the first dimension and range of all possible degrees (Theta) in the second dimension. In respect of a circle we can have 360 degrees possible. However not all degrees will be used in the calculation as our focus will mainly be on the horizontal and vertical lines. For that reason only certain degrees will be taken into consideration, such as 0 to 3 degrees, 90±3 degrees, 180±3 degrees, 270±3 degrees, and 357 to 360 degrees. To prevent the possibility of skew in the images, it is necessary for us to increase the calculation range by ±3 degrees. In addition by decreasing the range of Theta we increase the performance as the required calculation is reduced.

Geometrically horizontal line on an image can be represented by 90 and 270 degrees, and a vertical line by 0, 180 and 360 degrees. For instance, if we turn a 90 degree horizontal line clockwise for 180 degree it will still be a horizontal line, but its Theta has changed to 270.

Once we have an accumulator we then calculate “r” distance for each black pixel in opposition to every predefined degree, by using the formula mentioned above:

$$P = x \cos \vartheta + y \sin \vartheta$$

Where “P” represents the distance “r”, if the resulting distance is within the range of rMax, we then increment the accumulator value of the corresponding distance and degrees. Location of the pixel will also be stored for the purpose of line construction later on. However, a vital problem exists for us to recognise the line, as a line can be easily defined by two black pixels on the same axis. This could cause an unwanted line to be extracted from an image, because a line with many gaps can be formed through a word, sentence or paragraph. In order to resolve such problems and extract an authentic line from the image, a technique has been introduced to our algorithm.

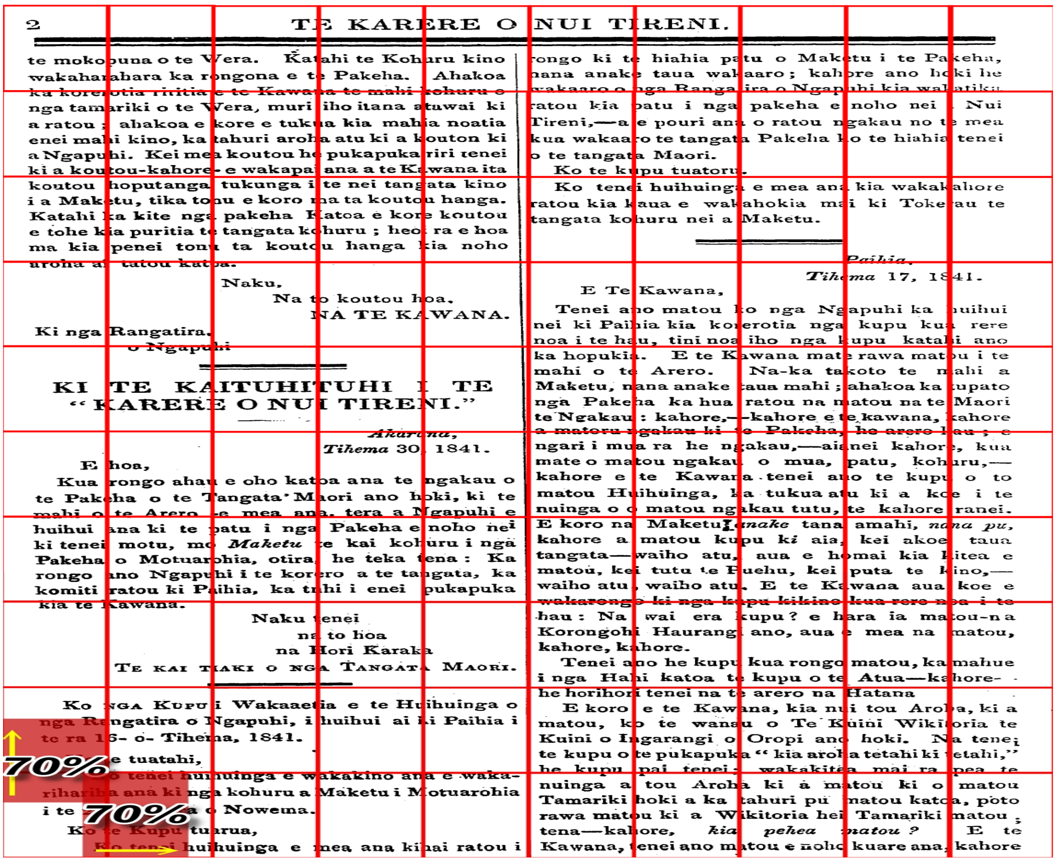
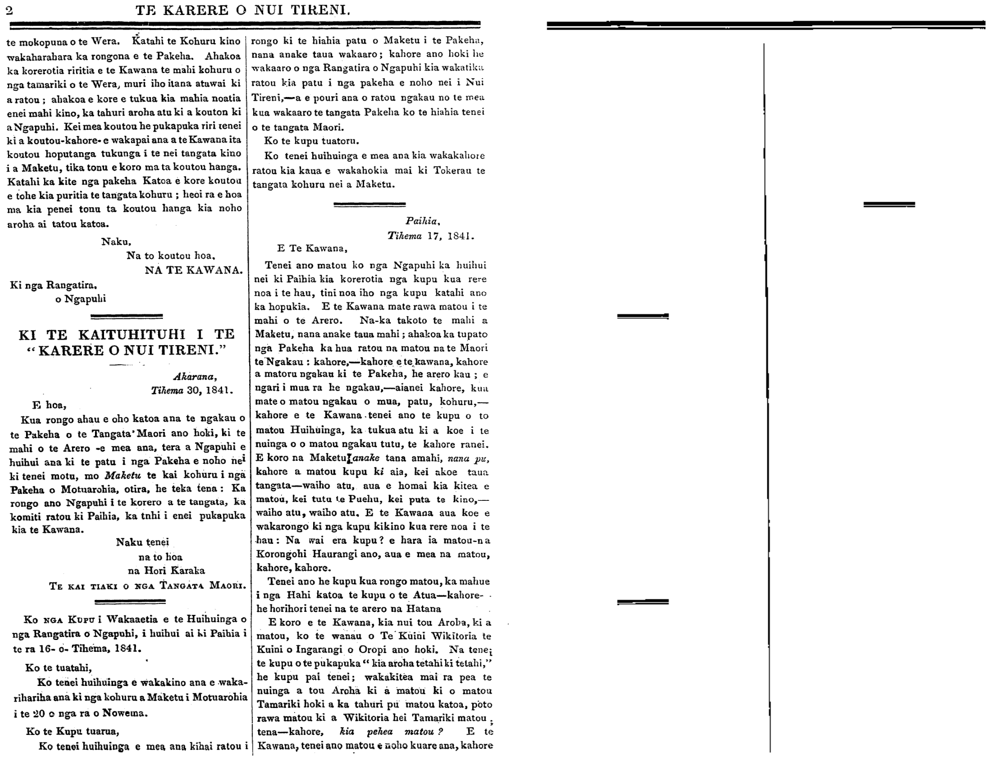


Figure 3.2a: Technique for authentic line detection.

The technique is shown in Figure 3.2a. The target image is divided into one hundred rectangular blocks, each with dimensions of one tenth to the width and height of the image. Hough Transform is then performed on one block at time. The threshold of 0.95 for the horizontal line and 0.85 for the vertical line, is then assigned to each block. Due to the nature of a line, we extract the horizontal line if the accumulator value reaches at least 95% of the block width, and 85% of the block height.

However, if we only apply the algorithm over one hundred blocks, segments of the lines might not be recognised as it could just be missed by the block. In preventing such a problem, a shift of the block is introduced. After each calculation a block is shifted 70% upwards and once it reaches the top the block is then placed to the bottom with 70% shifts to the right hand side. The whole process is then repeated until the last block reaches the end of the image. Extracted line segments can then be assembled together to form a logical representation of the line. In addition to this with the original line pixels recorded, it can then be used in the line removal algorithm later on.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atawai ki a ratou ; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru ; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena : Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

Ko NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro ; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

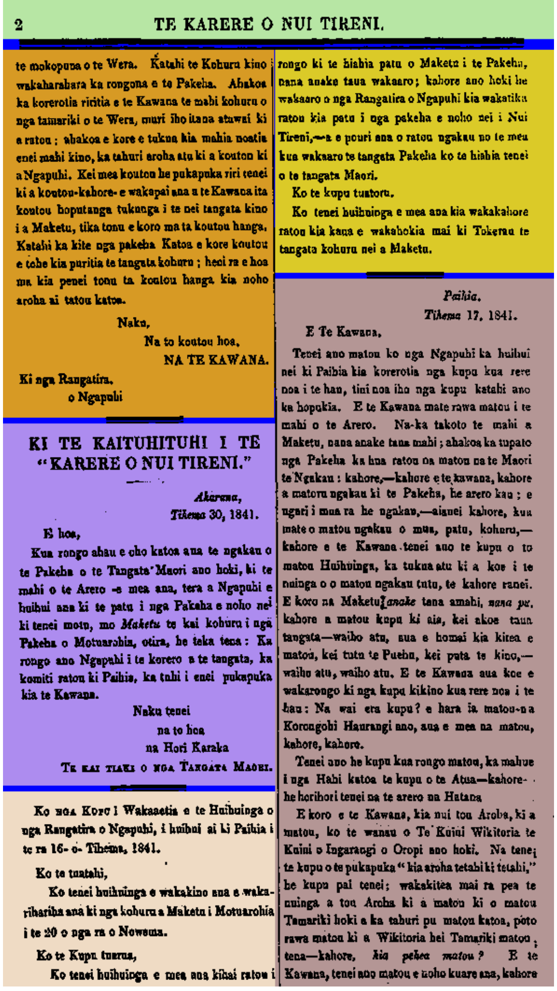
Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi ; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau : kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau ; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu *anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau : Na wai era kupu ? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroba, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka " kia aroha tetahi ki tetahi," he kupu pai tenei ; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou ; tena—kahore, *kia pehea matou ?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 3.2b: Results of the line extraction with Hough Transform (left: original image, right: line Image).

Figure 3.2b demonstrates an example of line extraction after applying Hough Transform to a document image. The horizontal and vertical lines are extracted, once the physical location of the line is identified, and the logical structure of the document can then be built using the extracted lines.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atawai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a koutou ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata' Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i ngä Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tuhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka

TE KAI TIAKI O NGA TANGATA MAORI.

Ko NGA KORO I Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake tana mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu *anake* tana amahi, *nana pu*, kahore a matou kupu ki aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Pueha, kei pata te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na a Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

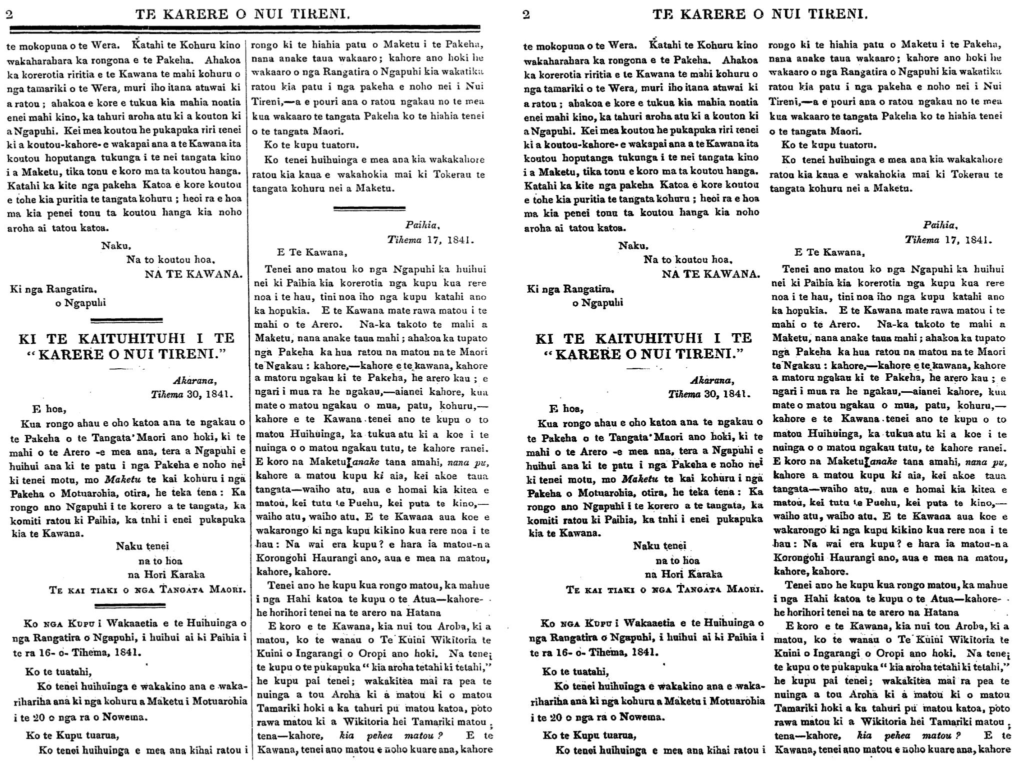
Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroha, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou; tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 3.2c: Results of the document segmentation using Hough Transform.

Figure 3.2c exemplifies an example of document segmentation based on lines extracted from the image. A robust logic core has been implemented in the algorithm, which allows it to determine and examine the relationships between each line (both horizontal and vertical), in order to locate an authentic position for segmenting the image into different sections. As shown above, the image has been segmented into six sections. Each represented with a different colour. As a result, logical structure of the document has been extracted. By using this information, layout of the document can then be reconstructed in the manner of paragraphs or columns.

During the implementation, there are number of times that drawing of the lines is required, in order to represent the extraction of the physical line. Bresenham's line algorithm (Bresenham 1965) has been employed in order to achieve an efficient line drawing within the algorithm. This algorithm was developed by J. E. Bresenham in 1962 at IBM. It is one of the earliest developed algorithms for computer graphics. Because of its simplicity and efficiency, it has been utilised in many modern graphic card chips and graphic libraries.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atawai ki a ratou ; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru ; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena : Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
Te kai tiaki o nga Tangata Maori.

Ko nga Kupu i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi ; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau : kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau ; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu[*anake* tana amahi, *nana pu,* kahore a matou kupu k*i* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau : Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroba, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou ; tena—kahore, *kia pehea matou ?* E te Kawana, tenei ano matou e noho kuare ana, kahore

2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atawai ki a ratou ; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru ; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena : Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
Te kai tiaki o nga Tangata Maori.

Ko nga Kupu i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi ; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau : kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau ; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu[*anake* tana amahi, *nana pu,* kahore a matou kupu k*i* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau : Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroba, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou ; tena—kahore, *kia pehea matou ?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 3.2d: Resulting images after applying line removal algorithm (left: before, right: after).

In addition an algorithm has been used based the on the Hough Transform for removing lines from the image. As it is shown in Figure 3.2b, lines can already be extracted from an image. However, numerous pixels connected to the line remained on the image, due to the fact that it did not match with the line extraction equation. It could be part of the noises or ink residues. Hence, in the algorithm, each line has been examined for any connected black pixels. Recursive process is utilised for complete noise removal around the line.

Figure 3.2d demonstrates the results, after applying the line removal algorithms to an image. This algorithm provides us with an automatic and efficient solution for absolute line removal from document images.

One of the drawbacks of Hough Transform is the costs of the calculation needed. Depending on the dimension of the image, it can be extremely time consuming if we perform Hough Transform on a large scale. As mentioned above, performance has already been improved by limiting the degrees required in the calculation. Furthermore, performance is improved by utilising an image scale down technique.

When an image is scaled down (made into a smaller size) it will start to lose pixels depending on the scaled down ratio. Nevertheless, structure of the line will still remain, even with a slightly distortion in an image. By using this characterises, the image has been scaled down into smaller dimension before Hough Transform, with scale down ratios (25%, 50%, 75%) based on the original dimension. Performance is then enhanced according to the amount of calculation reduced. Following this, segmentation results are then projected back to the original image based on the reduced ratio.

Initially Hough Transform is implemented using Lua script. However this is extremely inefficient. In order for an image with the dimensions of 1141 x 2043 to be fully segmented it requires a time of about sixteen minutes. As a result, the implementation has been changed directly using C++, and the time spent is reduced to approximately 8 seconds, which is about 120 times more efficient.

3.3 Corpus Used

In order to evaluate the implemented algorithms a corpus of historical newspapers was needed. In related projects mentioned in Chapter 2, there is some focus in using one newspaper collection for building a system, while others aim to build a system for archiving several newspaper collections. For example, on the smaller scale, The First International Newspaper Segmentation Contest (Gatos et al. 2001) used newspapers from both contemporary Greek and English, in addition to selecting the pages from different year publications, for the purpose of developing and evaluating the algorithms.

The most popular corpus for machine-printed document analysis (technical journals, etc) and OCR software analysis is by Phillips et al. (1993). It has been widely used or referred to in the research community. The corpus is published by the University of Washington, and consists of a database split over 3 CD-ROM discs. It cost around US $2 million to collate (Phillips et al. 1993). For our purposes, this database is ill-suited, however. Consequently we took the decision to assemble our own corpus. It is available as a DVD-ROM with this thesis. The appendix summarises its basic characteristics.

Newspaper images were gathered from twenty online newspaper collections and archives, such as Australian Newspapers, California Digital Newspaper Collection, Florida Digital Newspaper Library, Digital Archives Initiative, Digital Library of the Caribbean and Papers Past. Newspapers acquired from those archives are all in the public domain.

Five pages of newspapers have been gathered from each online collection for the purpose of developing and evaluating the implemented algorithms. Nevertheless, due to the inconsistence of the image quality, some images will appear to be of better quality than others. Each online newspaper archive might provide users with a different format of an historical newspaper. The majority of archives provide a PDF format of the newspaper while a few provided images in PNG or JPEG. For those newspapers with PDF format, raster images of the newspaper are extracted for the purpose of algorithm evaluation experiments.

The majority of newspapers are digitised with reasonable resolution i.e. characters are recognisable by human eyes. However there are a few newspapers with slightly lower resolution that are difficult to recognise. Newspaper images have a resolution ranging from 72 dpi to 300 dpi, as they are gathered from different sources. It is important to have a satisfactory amount of details within the image, in order to get sufficient segmentation and OCR. In addition all newspapers are within the range of the years 1800 to 1950.

Chapter 4 - Experiments

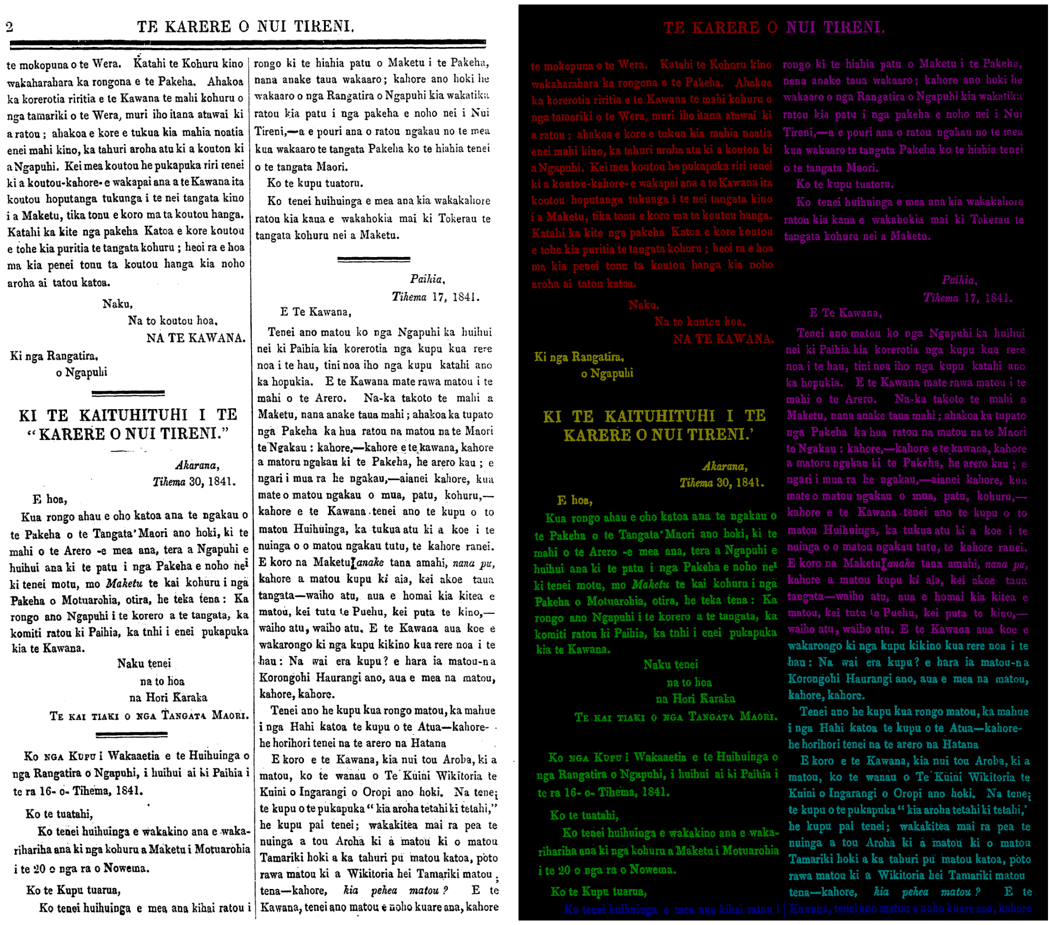
In this chapter, the basic algorithms applied to sample newspaper images are illustrated. Next the results of four major experiments presented, based on the compiled corpus in order to investigate the strengths and weakness of each page layout segmentation algorithm. This culminates in additional exploration of various combinations of different algorithms.

As mentioned before, Windows XP and Cygwin have been used for the installation of *Ocropus*. Additionally the experiments have been carried out on a laptop with AMD Turion64 ML-37 Mobile processer (2GHz), 2 Gigabytes of physical memory. The specifications of this laptop are not the most advantageous at the time of writing this thesis. However, it was sufficient for carrying out these experiments. We recommended that at least 2 Gigabyte of physical memory is available, in order to process larger scale images, which can consume significant amounts of memory depending on the algorithm used.

Experiments are performed based on three major algorithms: RAST, XY-Cut and Hough Transform, based on different historical newspaper layouts that were gathered from different online archives. We begin with simple layout and progress towards more complex ones. Experiments have also been performed on synthetic images using RAST and XY-Cut. Evaluation of performance with scaled down images using the Hough Transform and RAST algorithms is also presented. Finally results of applying a combination of algorithms on more complex historical newspaper layout are presented.

4.1 RAST Segmentation

RAST based layout analysis is one of the segmentation algorithms currently implemented in Ocropus. RAST algorithm is presently based on two other algorithms: whitespace identification and constrained text line finding (Breuel and Kaiserlautern 2007). In other words, whitespaces between each component and existing evidence of text lines are both imperative factors during the RAST segmentation.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

Ko NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu[*anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroba, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou; tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE KARERE O NUI TIRENI.'

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

Ko NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu[*anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroba, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi,' he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

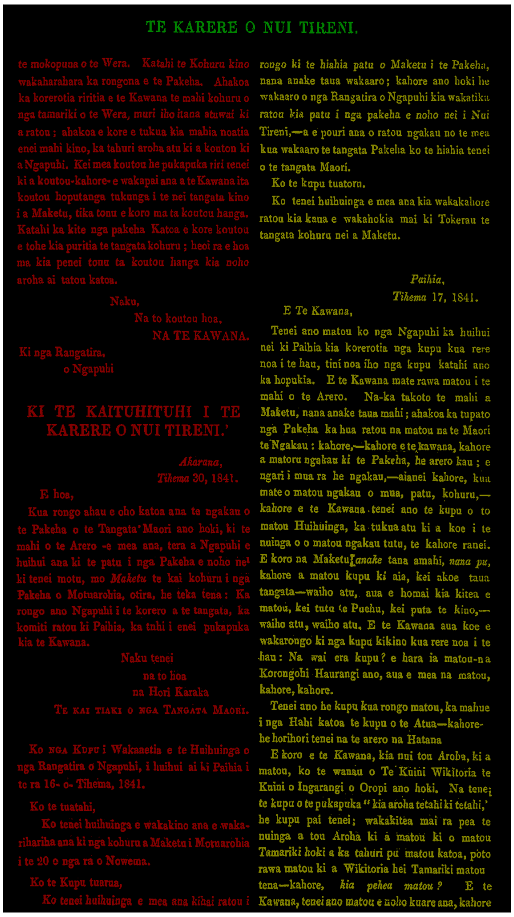
Figure 4.1: Example of RAST segmentation (column group, image dimension: 1141 x 2043, resolution: 96 DPI).

Figure 4.1 illustrates an image after segmentation by the RAST algorithm, followed by grouping the components into columns (on this particular image it does not perform well). The image is acquired from Maori newspaper collections (Niupepa) hosted in New Zealand Digital Library Website (NZDL 2009). RAST segmented the image into different components. The column grouping function, provided in *Ocropus*, is then applied to the segmented result which presents each column in a different colour. As it shows, RAST ignores the non-text components or components that do not fall in a line of text. Both lines and the page number on the top left hand corner have been omitted by the algorithm. Furthermore, it separates the title into two segments but fails to separate the last line of text into two segments. Paragraphs in each column have also been unsuccessful in being merged into one column.

There could be many reasons for imperfect segmentation. One of the most likely reasons is the existence of the skew, although to the naked eye it can be difficult to discern. In order to determine the level of skew in the image, the skew estimation function, was applied to the image. As the result, the image is estimated with approximately 0.52 degrees of skew.

Using this information, the image was corrected for skew and re-processed using the RAST segmentation. As is shown in Figure 4.2, the image is now perfectly segmented into a title and two separate columns.

This provides us with evidence of how sensitive the RAST algorithm is to the existence of skew. Even with very little amount of skew (less than one degree), it can still have a significant effect on the segmentation results.



TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE KARERE O NUI TIRENI.'

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tuhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

KO NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou *kia* patu i nga pakeha e *noho* nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

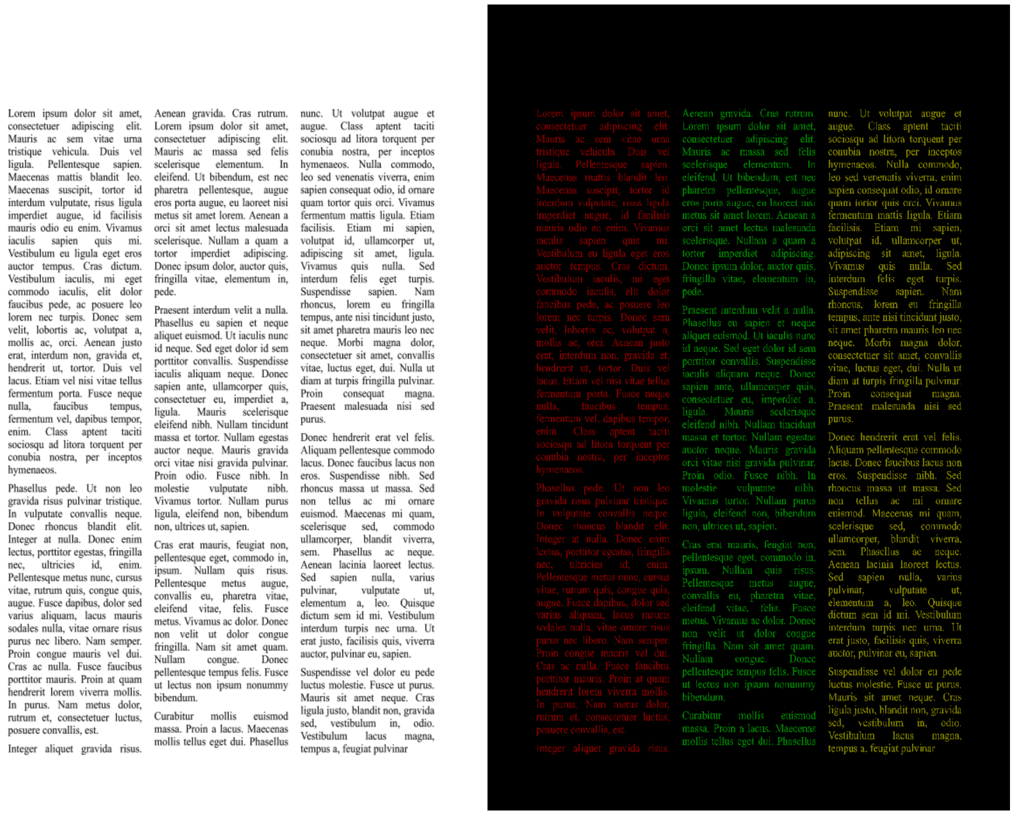
Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau *ki* te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu[*anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, *kei* tutu te Puehu, *kei* puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore-he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroba, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka " kia aroha tetahi ki tetahi,' he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 4.2: Example of RAST segmentation after de-skew (column group).

In order to further explore the relationships between skew and the RAST algorithm, a computer generated synthetic image (using Microsoft Word) was created to be tested by RAST. There is a major difference between historical newspapers and synthetic documents as such, as they are typed directly on to the computer. Therefore, each character is guaranteed to be printed accurately. In addition, the document is free from any skews and noises. Spaces between each character, paragraph and column are perfectly consistent.



Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Mauris ac sem vitae urna tristique vehicula. Duis vel ligula. Pellentesque sapien. Maecenas mattis blandit leo. Maecenas suscipit, tortor id interdum vulputate, risus ligula imperdiet augue, id facilisis mauris odio eu enim. Vivamus iaculis sapien quis mi. Vestibulum eu ligula eget eros auctor tempus. Cras dictum. Vestibulum iaculis, mi eget commodo iaculis, elit dolor faucibus pede, ac posuere leo lorem nec turpis. Donec sem velit, lobortis ac, volutpat a, mollis ac, orci. Aenean justo erat, interdum non, gravida et, hendrerit ut, tortor. Duis vel lacus. Etiam vel nisi vitae tellus fermentum porta. Fusce neque nulla, faucibus tempus, fermentum vel, dapibus tempor, enim. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos.

Phasellus pede. Ut non leo gravida risus pulvinar tristique. In vulputate convallis neque. Donec rhoncus blandit elit. Integer at nulla. Donec enim lectus, porttitor egestas, fringilla nec, ultricies id, enim. Pellentesque metus nunc, cursus vitae, rutrum quis, congue quis, augue. Fusce dapibus, dolor sed varius aliquam, lacus mauris sodales nulla, vitae ornare risus purus nec libero. Nam semper. Proin congue mauris vel dui. Cras ac nulla. Fusce faucibus porttitor mauris. Proin at quam hendrerit lorem viverra mollis. In purus. Nam metus dolor, rutrum et, consectetuer luctus, posuere convallis, est.

Integer aliquet gravida risus. Aenean gravida. Cras rutrum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Mauris ac massa sed felis scelerisque elementum. In eleifend. Ut bibendum, est nec pharetra pellentesque, augue eros porta augue, eu laoreet nisi metus sit amet lorem. Aenean a orci sit amet lectus malesuada scelerisque. Nullam a quam a tortor imperdiet adipiscing. Donec ipsum dolor, auctor quis, fringilla vitae, elementum in, pede.

Praesent interdum velit a nulla. Phasellus eu sapien et neque aliquet euismod. Ut iaculis nunc id neque. Sed eget dolor id sem porttitor convallis. Suspendisse iaculis aliquam neque. Donec sapien ante, ullamcorper quis, consectetuer eu, imperdiet a, ligula. Mauris scelerisque eleifend nibh. Nullam tincidunt massa et tortor. Nullam egestas auctor neque. Mauris gravida orci vitae nisi gravida pulvinar. Proin odio. Fusce nibh. In molestie vulputate nibh. Vivamus tortor. Nullam purus ligula, eleifend non, bibendum non, ultrices ut, sapien.

Cras erat mauris, feugiat non, pellentesque eget, commodo in, ipsum. Nullam quis risus. Pellentesque metus augue, convallis eu, pharetra vitae, eleifend vitae, felis. Fusce metus. Vivamus ac dolor. Donec non velit ut dolor congue fringilla. Nam sit amet quam. Nullam congue. Donec pellentesque tempus felis. Fusce ut lectus non ipsum nonummy bibendum.

Curabitur mollis euismod massa. Proin a lacus. Maecenas mollis tellus eget dui. Phasellus nunc. Ut volutpat augue et augue. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Nulla commodo, leo sed venenatis viverra, enim sapien consequat odio, id ornare quam tortor quis orci. Vivamus fermentum mattis ligula. Etiam facilisis. Etiam mi sapien, volutpat id, ullamcorper ut, adipiscing sit amet, ligula. Vivamus quis nulla. Sed interdum felis eget turpis. Suspendisse sapien. Nam rhoncus, lorem eu fringilla tempus, ante nisi tincidunt justo, sit amet pharetra mauris leo nec neque. Morbi magna dolor, consectetuer sit amet, convallis vitae, luctus eget, dui. Nulla ut diam at turpis fringilla pulvinar. Proin consequat magna. Praesent malesuada nisi sed purus.

Donec hendrerit erat vel felis. Aliquam pellentesque commodo lacus. Donec faucibus lacus non eros. Suspendisse nibh. Sed rhoncus massa ut massa. Sed non tellus ac mi ornare euismod. Maecenas mi quam, scelerisque sed, commodo ullamcorper, blandit viverra, sem. Phasellus ac neque. Aenean lacinia laoreet lectus. Sed sapien nulla, varius pulvinar, vulputate ut, elementum a, leo. Quisque dictum sem id mi. Vestibulum interdum turpis nec urna. Ut erat justo, facilisis quis, viverra auctor, pulvinar eu, sapien.

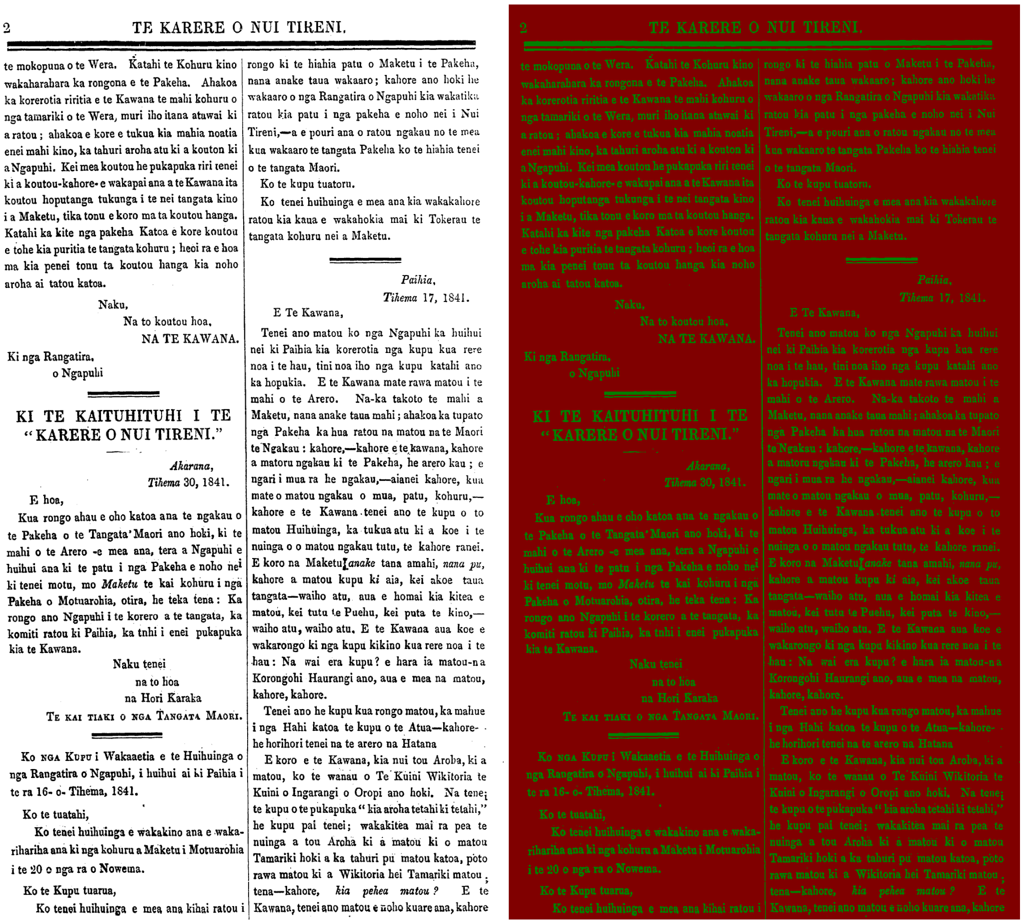
Suspendisse vel dolor eu pede luctus molestie. Fusce ut purus. Mauris sit amet neque. Cras ligula justo, blandit non, gravida sed, vestibulum in, odio. Vestibulum lacus magna, tempus a, feugiat pulvinar

Figure 4.3: Example of RAST segmentation on synthetic image (column group, image dimension: 1700 x 2339, resolution: 96 DPI).

As shown in Figure 4.3, a three column synthetic document image has been generated. Without prior instruction to de-skew, the image is directly segmented by RAST algorithm. As a result, the image is perfectly segmented into three columns. Again, it provides us with evidence that the RAST algorithm works well when no skew is present. The algorithm is well suited for segmenting documents with Manhattan type layout (journals and so forth), where the components within the image are perfectly aligned. The segmentation test was repeated with synthetic images of four and five columns with comparable results.

4.2 XY-Cut Segmentation

A prototype of the XY-Cut segmentation algorithm has also been implemented in Ocropus recently. As mentioned above, this algorithm is implemented using the tree structure, followed by segmenting the image based on horizontal and vertical gaps between each component.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata'Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

Ko NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu *anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore-he horihori tenei na te arero na Hatana

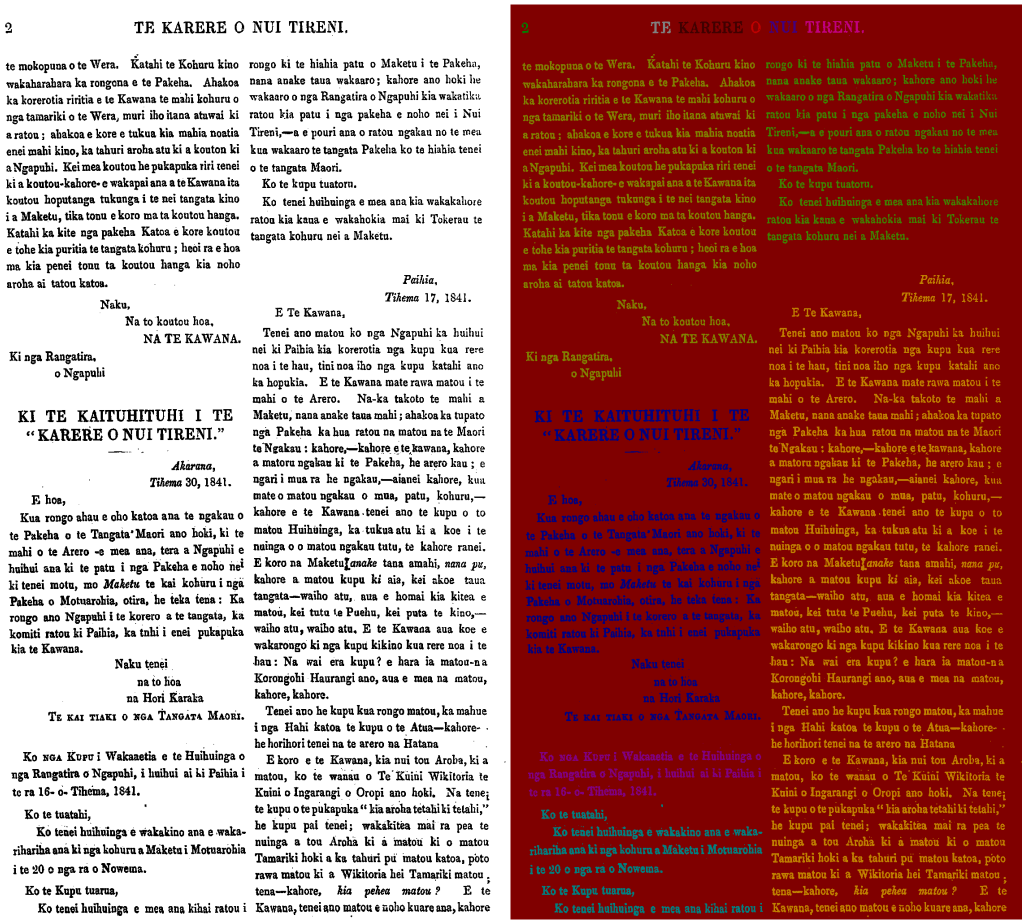
E koro e te Kawana, kia nui tou Aroha, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou, tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 4.4: Example of using XY-Cut segmentation (image dimension: 1141 x 2043, resolution: 96 DPI).

In order to evaluate the use of XY-cut algorithm on historical newspapers, we applied it to the same image used in the RAST algorithm. As is shown in Figure

4.4, the image on the right hand side, shows the XY-Cut algorithm has some difficulty segmenting the images into different sections. Every component on the page has been segmented as one section.

We hypothesised that the reason for incorrect segmentation was caused by interference of the ruled lines. Hence, lines in the image were removed by using the Hough Transform developed line removal algorithm. We then fed the processed image to XY-Cut algorithm again.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata' Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

KO NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu *anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore-he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroha, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou, tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita koutou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata' Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena: Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

KO NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakarihariha ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i rongo ki te hiahia patu o Maketu i te Pakeha, nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatika ratou kia patu i nga pakeha e noho nei i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

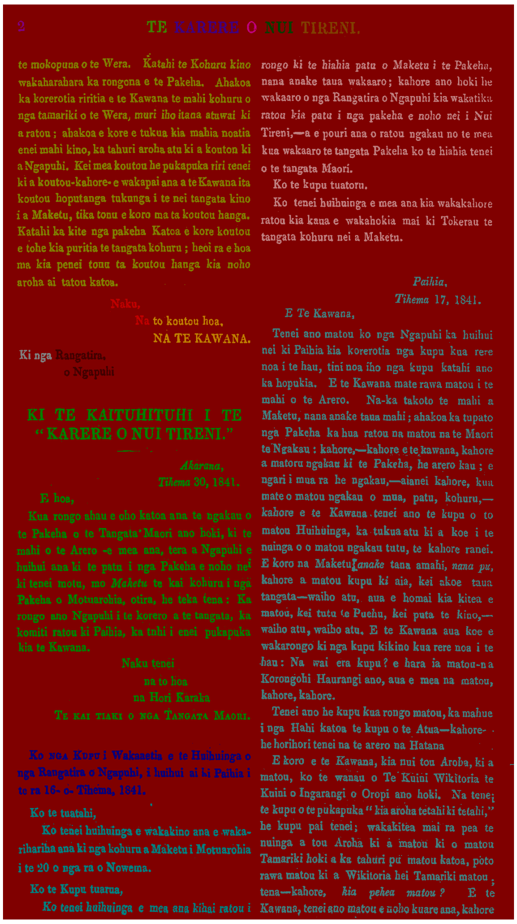
Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau: kahore,—kahore e te kawana, kahore a matoru ngakau ki te Pakeha, he arero kau; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu *anake* tana amahi, *nana pu*, kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau: Na wai era kupu? e hara ia matou-na Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore-he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroha, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tenei te kupu o te pukapuka "kia aroha tetahi ki tetahi," he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou, tena—kahore, *kia pehea matou?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 4.5: XY-Cut segmentation after line removed from an image.

The results shown in Figure 4.5, the XY-Cut algorithm succeeds in segmenting the image into different sections. Compared to RAST segmentation it did not omit the page number at the top left corner. It also has, in this example, more accurate detection for paragraphs. This could be due to its nature of segmenting components based on gaps between each element. However, it failed to group the image title into one section, due to the fact that the title was printed with wider gaps, compared with sentences in the paragraph.



2 TE KARERE O NUI TIRENI.

te mokopuna o te Wera. Katahi te Kohuru kino wakaharahara ka rongona e te Pakeha. Ahakoa ka korerotia riritia e te Kawana te mahi kohuru o nga tamariki o te Wera, muri iho itana atuwai ki a ratou ; ahakoa e kore e tukua kia mahia noatia enei mahi kino, ka tahuri aroha atu ki a kouton ki a Ngapuhi. Kei mea koutou he pukapuka riri tenei ki a koutou-kahore- e wakapai ana a te Kawana ita kontou hoputanga tukunga i te nei tangata kino i a Maketu, tika tonu e koro ma ta koutou hanga. Katahi ka kite nga pakeha Katoa e kore koutou e tohe kia puritia te tangata kohuru ; heoi ra e hoa ma kia penei tonu ta koutou hanga kia noho aroha ai tatou katoa.

Naku,
Na to koutou hoa,
NA TE KAWANA.

Ki nga Rangatira,
o Ngapuhi

KI TE KAITUHITUHI I TE "KARERE O NUI TIRENI."

Akarana,
Tihema 30, 1841.

E hoa,

Kua rongo ahau e oho katoa ana te ngakau o te Pakeha o te Tangata' Maori ano hoki, ki te mahi o te Arero -e mea ana, tera a Ngapuhi e huihui ana ki te patu i nga Pakeha e noho nei ki tenei motu, mo *Maketu* te kai kohuru i nga Pakeha o Motuarohia, otira, he teka tena : Ka rongo ano Ngapuhi i te korero a te tangata, ka komiti ratou ki Paihia, ka tnhi i enei pukapuka kia te Kawana.

Naku tenei
na to hoa
na Hori Karaka
TE KAI TIAKI O NGA TANGATA MAORI.

Ko NGA KUPU i Wakaaetia e te Huihuinga o nga Rangatira o Ngapuhi, i huihui ai ki Paihia i te ra 16- o- Tihema, 1841.

Ko te tuatahi,

Ko tenei huihuinga e wakakino ana e wakariharihа ana ki nga kohuru a Maketu i Motuarohia i te 20 o nga ra o Nowema.

Ko te Kupu tuarua,

Ko tenei huihuinga e mea ana kihai ratou i *rongo ki te hiahia patu o Maketu i te Pakeha,* nana anake taua wakaaro; kahore ano hoki he wakaaro o nga Rangatira o Ngapuhi kia wakatiku *ratou kia patu i nga pakeha e noho nei* i Nui Tireni,—a e pouri ana o ratou ngakau no te mea kua wakaaro te tangata Pakeha ko te hiahia tenei o te tangata Maori.

Ko te kupu tuatoru.

Ko tenei huihuinga e mea ana kia wakakahore ratou kia kaua e wakahokia mai ki Tokerau te tangata kohuru nei a Maketu.

Paihia,
Tihema 17, 1841.

E Te Kawana,

Tenei ano matou ko nga Ngapuhi ka huihui nei ki Paihia kia korerotia nga kupu kua rere noa i te hau, tini noa iho nga kupu katahi ano ka hopukia. E te Kawana mate rawa matou i te mahi o te Arero. Na-ka takoto te mahi a Maketu, nana anake taua mahi ; ahakoa ka tupato nga Pakeha ka hua ratou na matou na te Maori te Ngakau : kahore,—kahore e te kawana, kahore a matoru ngakau *ki* te *Pakeha*, he arero kau ; e ngari i mua ra he ngakau,—aianei kahore, kua mate o matou ngakau o mua, patu, kohuru,—kahore e te Kawana tenei ano te kupu o to matou Huihuinga, ka tukua atu ki a koe i te nuinga o o matou ngakau tutu, te kahore ranei. E koro na Maketu[*anake* tana amahi, *nana pu,* kahore a matou kupu *ki* aia, kei akoe taua tangata—waiho atu, aua e homai kia kitea e matou, kei tutu te Puehu, kei puta te kino,—waiho atu, waiho atu. E te Kawana aua koe e wakarongo ki nga kupu kikino kua rere noa i te hau : Na wai era kupu ? e hara ia matou-na a Korongohi Haurangi ano, aua e mea na matou, kahore, kahore.

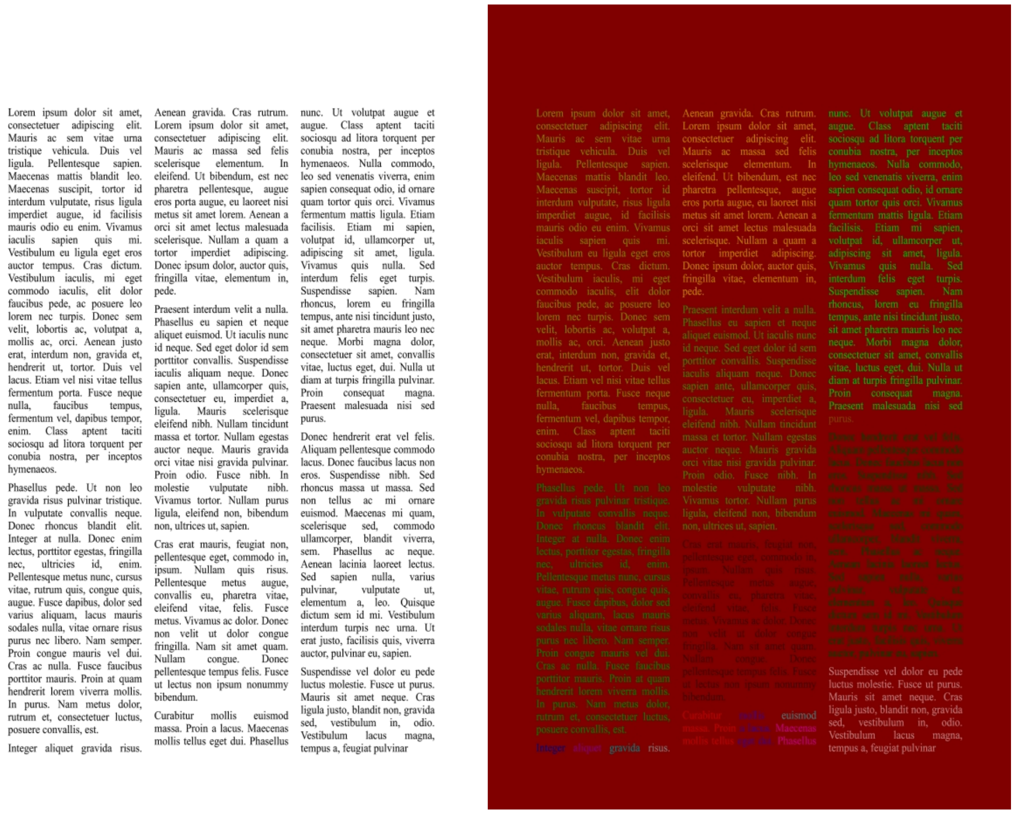
Tenei ano he kupu kua rongo matou, ka mahue i nga Hahi katoa te kupu o te Atua—kahore- he horihori tenei na te arero na Hatana

E koro e te Kawana, kia nui tou Aroha, ki a matou, ko te wanau o Te Kuini Wikitoria te Kuini o Ingarangi o Oropi ano hoki. Na tene; te kupu o te pukapuka "*kia aroha tetahi ki tetahi,*" he kupu pai tenei; wakakitea mai ra pea te nuinga a tou Aroha ki a matou ki o matou Tamariki hoki a ka tahuri pu matou katoa, poto rawa matou ki a Wikitoria hei Tamariki matou ; tena—kahore, *kia pehea matou ?* E te Kawana, tenei ano matou e noho kuare ana, kahore

Figure 4.6: XY-Cut segmentation after de-skew the image.

Unexpected segmentation results to some degree, can be formed after we apply de-skew to the image. Figure 4.6 illustrates the segmentation results after the image has been de-skewed. Few extra sections have been formed under the first paragraph. This is caused by slight changes to the gaps between each component during the de-skewing process.

Nevertheless, a point worth mentioning is the performance of the XY-Cut algorithm. The image in Figure 4.6 is fully segmented in approximately 1.3 seconds, compared with Hough Transform (described next) taking approximately 9 seconds. For XY-Cut, time spent is based on both dimensions and different document layouts, but generally speaking it is more efficient than Hough.



Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Mauris ac sem vitae urna tristique vehicula. Duis vel ligula. Pellentesque sapien. Maecenas mattis blandit leo. Maecenas suscipit, tortor id interdum vulputate, risus ligula imperdiet augue, id facilisis mauris odio eu enim. Vivamus iaculis sapien quis mi. Vestibulum eu ligula eget eros auctor tempus. Cras dictum. Vestibulum iaculis, mi eget commodo iaculis, elit dolor faucibus pede, ac posuere leo lorem nec turpis. Donec sem velit, lobortis ac, volutpat a, mollis ac, orci. Aenean justo erat, interdum non, gravida et, hendrerit ut, tortor. Duis vel lacus. Etiam vel nisi vitae tellus fermentum porta. Fusce neque nulla, faucibus tempus, fermentum vel, dapibus tempor, enim. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos.

Phasellus pede. Ut non leo gravida risus pulvinar tristique. In vulputate convallis neque. Donec rhoncus blandit elit. Integer at nulla. Donec enim lectus, porttitor egestas, fringilla nec, ultricies id, enim. Pellentesque metus nunc, cursus vitae, rutrum quis, congue quis, augue. Fusce dapibus, dolor sed varius aliquam, lacus mauris sodales nulla, vitae ornare risus purus nec libero. Nam semper. Proin congue mauris vel dui. Cras ac nulla. Fusce faucibus porttitor mauris. Proin at quam hendrerit lorem viverra mollis. In purus. Nam metus dolor, rutrum et, consectetuer luctus, posuere convallis, est.

Integer aliquet gravida risus. Aenean gravida. Cras rutrum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Mauris ac massa sed felis scelerisque elementum. In eleifend. Ut bibendum, est nec pharetra pellentesque, augue eros porta augue, eu laoreet nisi metus sit amet lorem. Aenean a orci sit amet lectus malesuada scelerisque. Nullam a quam a tortor imperdiet adipiscing. Donec ipsum dolor, auctor quis, fringilla vitae, elementum in, pede.

Praesent interdum velit a nulla. Phasellus eu sapien et neque aliquet euismod. Ut iaculis nunc id neque. Sed eget dolor id sem porttitor convallis. Suspendisse iaculis aliquam neque. Donec sapien ante, ullamcorper quis, consectetuer eu, imperdiet a, ligula. Mauris scelerisque eleifend nibh. Nullam tincidunt massa et tortor. Nullam egestas auctor neque. Mauris gravida orci vitae nisi gravida pulvinar. Proin odio. Fusce nibh. In molestie vulputate nibh. Vivamus tortor. Nullam purus ligula, eleifend non, bibendum non, ultrices ut, sapien.

Cras erat mauris, feugiat non, pellentesque eget, commodo in, ipsum. Nullam quis risus. Pellentesque metus augue, convallis eu, pharetra vitae, eleifend vitae, felis. Fusce metus. Vivamus ac dolor. Donec non velit ut dolor congue fringilla. Nam sit amet quam. Nullam congue. Donec pellentesque tempus felis. Fusce ut lectus non ipsum nonummy bibendum.

Curabitur mollis euismod massa. Proin a lacus. Maecenas mollis tellus eget dui. Phasellus nunc. Ut volutpat augue et augue. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Nulla commodo, leo sed venenatis viverra, enim sapien consequat odio, id ornare quam tortor quis orci. Vivamus fermentum mattis ligula. Etiam facilisis. Etiam mi sapien, volutpat id, ullamcorper ut, adipiscing sit amet, ligula. Vivamus quis nulla. Sed interdum felis eget turpis. Suspendisse sapien. Nam rhoncus, lorem eu fringilla tempus, ante nisi tincidunt justo, sit amet pharetra mauris leo nec neque. Morbi magna dolor, consectetuer sit amet, convallis vitae, luctus eget, dui. Nulla ut diam at turpis fringilla pulvinar. Proin consequat magna. Praesent malesuada nisi sed purus.

Donec hendrerit erat vel felis. Aliquam pellentesque commodo lacus. Donec faucibus lacus non eros. Suspendisse nibh. Sed rhoncus massa ut massa. Sed non tellus ac mi ornare euismod. Maecenas mi quam, scelerisque sed, commodo ullamcorper, blandit viverra, sem. Phasellus ac neque. Aenean lacinia laoreet lectus. Sed sapien nulla, varius pulvinar, vulputate ut, elementum a, leo. Quisque dictum sem id mi. Vestibulum interdum turpis nec urna. Ut erat justo, facilisis quis, viverra auctor, pulvinar eu, sapien.

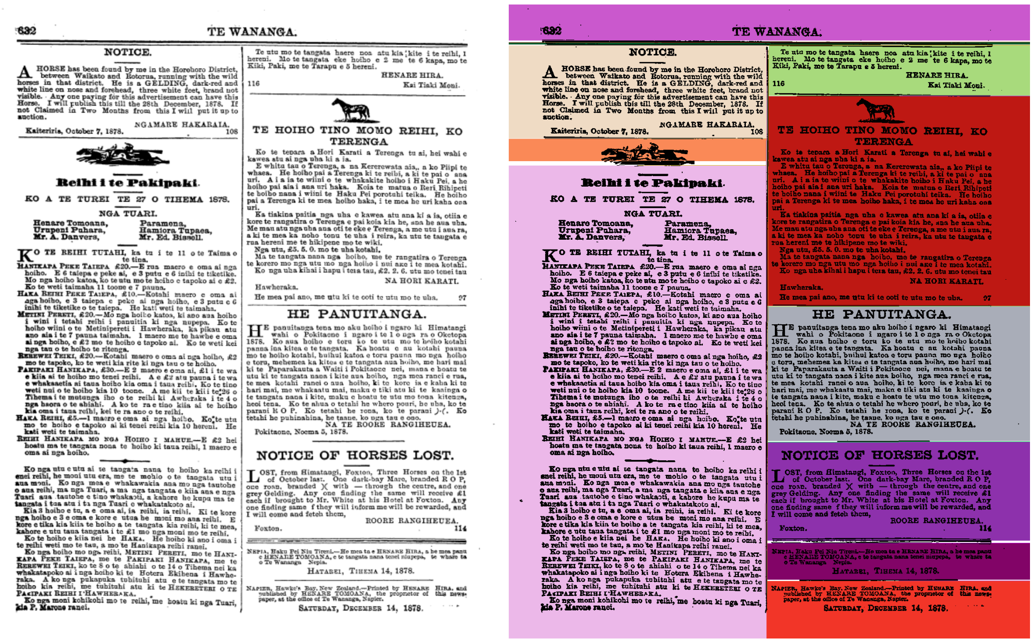
Suspendisse vel dolor eu pede luctus molestie. Fusce ut purus. Mauris sit amet neque. Cras ligula justo, blandit non, gravida sed, vestibulum in, odio. Vestibulum lacus magna, tempus a, feugiat pulvinar

Figure 4.7: Example of XY-Cut segmentation on synthetic image (image dimension: 1700 x 2339, resolution: 96 DPI).

Additionally, segmentation for synthetic documents have been tested with the XY-Cut algorithm. Using the same as was used previously with the RAST algorithm. As it is shown in Figure 4.7, most paragraphs have been successfully identified. However, the last few sentences at the end of first and second columns have been segmented into several different components.

Compared with RAST, the current implementation of the XY-Cut algorithm is less adequate in dealing with Manhattan-like layout documents. Further fine tuning to the implementation will be needed for the XY-Cut algorithm, in order for it to acquire better segmentation results.

4.3 Hough Transform Segmentation



632 TE WANANGA.

NOTICE.

A HORSE has been found by me in the Horohoro District, between Waikato and Rotorua, running with the wild horses in that district. He is a GELDING, dark-red and white line on nose and forehead, three white feet, brand not visible. Any one paying for this advertisement can have this Horse. I will publish this till the 28th December, 1878. If not Claimed in Two Months from this I will put it up to auction.

NGAMARE HAKARAIA.
Kaiteriria, October 7, 1878. 108

Reihi i te Pakipaki.

KO A TE TUREI TE 27 O TIHEMA 1878.

NGA TUARI.

Henare Tomoana,	Paramena,
Urupeni Puhara,	Hamiora Tupaea,
Mr. A. Danvers,	Mr. Ed. Bissell.

KO TE REIHI TUTAHI, ka tu i te 11 o te Taima o te tina.

Hanikapa Peke Taiepa £20.—E rua maero e oma ai nga hoiho. E 6 taiepa e peke ai, e 3 putu e 6 inihi te tiketike. Mo nga hoiho katoa, ko te utu mo te hoiho e tapoko ai e £2. Ko te weti taimaha 11 toone e 7 pauna.

Haka Reihi Peke Taiepa, £10.—Kotahi maero e oma ai nga hoiho, e 3 taiepa e peke ai nga hoiho, e 3 putu e 6 inihi te tiketike o te taiepa. He kati weti te taimaha.

Metini Pereti, £20.—Mo nga hoiho katoa, ki ano aua hoiho i wini i tetahi reihi i panuitia ki nga nupepa. Ko te hoiho wiini o te Metinipereti i Hawheraka, ka pikau atu ano aia i te 7 pauna taimaha. 1 maero me te hawhe e oma ai nga hoiho, e £2 mo te hoiho e tapoko ai. Ko te weti kei nga tau o te hoiho te ritenga.

Rerewei Teiki, £20.—Kotahi maero e oma ai nga hoiho, £2 mo te tapoko, ko te weti kia rite ki nga tau o te hoiho.

Pakipaki Hanikapa, £30.—E 2 maero e oma ai, £1 i te wa e kiia ai te hoiho mo tenei reihi. A e £2 atu pauna i te wa e whakaaetia ai taua hoiho kia oma i taua reihi. Ko te tino weti nui o te hoiho kia 10 toone. A me kii te kii i te 26 o Tihema i te mutunga iho o te reihi ki Awheraka i te 4 o nga haora o te ahiahi. A ko te ra e tino kiia ai te hoiho kia oma i taua reihi, kei te ra ano o te reihi.

Haka Reihi, £5.—1 maero e oma ai nga hoiho. Ko te utu mo te hoiho e tapoko ai ki tenei reihi kia 10 hereni. He kati weti te taimaha.

Reihi Hanikapa mo nga Hoiho i mahue.—E £2 hei hoatu ma te tangata nona te hoiho ki taua reihi, 1 maero e oma ai nga hoiho.

Ko nga utu e utu ai te tangata nana te hoiho ka reihi i enei reihi, he moni utu era, me te mohio o te tangata utu i aua moni. Ko nga mea e whakawakia ana mo nga tautohe o aua reihi, ma nga Tuari, a ma nga tangata e kiia ana e nga Tuari aua tautohe e tino whakaoti, a kahore he kupu ma te tangata i tua atu i ta nga Tuari e whakatakoto ai.

Kia 3 hoiho e tu, a e oma ai, ia reihi, ia reihi. Ki te kore nga hoiho e 3 e oma e kore e utua he moni mo aua reihi. E kore e tika kia kiia te hoiho a te tangata kia reihi, ki te mea, kahore e utu taua tangata i te £1 mo nga moni mo te reihi.

Ko te hoiho e kiia nei he Haka. He hoiho ki ano i oma i te reihi weti mo te tau, a mo te Hanikapa reihi ranei.

Ko nga hoiho mo nga reihi, Metini Pereti, mo te Hanikapa Peke Taiepa, me te Pakipaki Hanikapa, me te Rerewei Teiki, ko te 8 o te ahiahi o te 14 o Tihema nei ka whakatapoko ai i nga hoiho ki te Hotera Ekihena i Hawheraka. A ko nga pukapuka tuhituhi atu e te tangata mo te hoiho kia reihi, me tuhituhi atu ki te Hekereteri o te Pakipaki Reihi i Hawheraka.

Ko nga moni kohikohi mo te reihi, me hoatu ki nga Tuari, kia P. Marone ranei.

Te utu mo te tangata haere noa atu kia kite i te reihi, 1 hereni. Mo te tangata eke hoiho e 2 me te 6 kapa, mo te Kiki, Paki, me te Tarapu e 5 hereni.

HENARE HIRA.
116 Kai Tiaki Moni.

TE HOIHO TINO MOMO REIHI, KO TERENGA

Ko te tepara a Hori Karati a Terenga tu ai, hei wahi e kawea atu ai nga uha ki a ia.

E whitu tau o Terenga, a na Kererewata aia,, a ko Piipi te whaea. He hoiho pai a Terenga ki te reihi, a ki te pai o ana uri. A i a ia te wiini o te whakakite hoiho i Haku Pei, a he hoiho pai aia i ana uri haka. Koia te matua o Reri Rihipeti te hoiho nana i wiini te Haku Pei porotuhi teika. He hoiho pai a Terenga ki te mea hoiho haka, i te mea he uri kaha ona uri.

Ka tiakina paitia nga uha e kawea atu ana ki a ia, otiia e kore te rangatira o Terenga e pai koia kia he, ana he aua uha. Me mau atu nga uha ana oti te eke e Terenga, a me utu i aua ra, a ki te mea ka noho tonu te uha i reira, ka utu te tangata e rua hereni me te hikipene mo te wiki.

Nga utu, £5. 5. 0. mo te uha kotahi.

Ma te tangata nana nga hoiho, me te rangatira o Terenga te korero mo nga utu mo nga hoiho i nui ake i te mea kotahi.

Ko nga uha kihai i hapu i tera tau, £2. 2. 6. utu mo tenei tau

NA HORI KARATI.
Hawheraka.

He mea pai ano, me utu ki te coti te utu mo te uha. 97

HE PANUITANGA.

HE panuitanga tena mo aku hoiho i ngaro ki Himatangi wahi o Pokitaone i ngaro i te 1 o nga ra o Oketopa 1878. Ko aua hoiho e toru ko te utu mo te hoiho kotahi pauna ina kitea e te tangata. Ka hoatu e au kotahi pauna mo te hoiho kotahi, huihui katoa e toru pauna mo nga hoiho e toru, mehemea ka kitea e te tangata aua hoiho, me hari mai ki te Paparakauta a Waiti i Pokitaone nei, mana e hoatu te utu ki te tangata nana i kite aua hoiho, nga mea ranei e rua, te mea kotahi ranei o aua hoiho, ki te kore ia e kaha ki te hari mai, me whakaatu mai, maku e tiki atu ki te kainga o te tangata nana i kite, maku e hoatu te utu mo tona kitenga, heoi tena. Ko te ahua o tetahi he whero pouri, he uha, ko te parani R O P. Ko tetahi he rona, ko te parani)-(. Ko tetahi he puhinahina, he taane. ko nga tau e ono.

NA TE ROORE RANGIHEUEA.
Pokitaone, Noema 5, 1878.

NOTICE OF HORSES LOST.

LOST, from Himatangi, Foxton, Three Horses on the 1st of October last. One dark-bay Mare, branded R O P, one roan, branded)(with — through the centre, and one grey Gelding. Any one finding the same will receive £1 each if brought to Mr. White at his Hotel at Foxton. Any one finding same f they will inform me will be rewarded, and I will come and fetch them,

ROORE RANGIHEUEA.
Foxton. 114

Nepia, Haku Pei Niu Tireni.—He mea ta e Henare Hira, a he mea panu e HENARE TOMOANA, e te tangata nana tenei niupepa, te whare ta o Te Wananga Nepia.

Hatarei, Tihema 14, 1878.

Napier, Hawke's Bay, New Zealand.—Printed by Henare Hira, and published by HENARE TOMOANA, the proprietor of this newspaper, at the office of Te Wananga, Napier.

Saturday, December 14, 1878.

Figure 4.8: Example of Hough Transform based segmentation (image dimension: 1582 x 1928, resolution: 96 DPI).

As mentioned previously, for this thesis an implementation of the Hough Transform was added into *Ocropus* for the purpose of segmenting historical newspapers. Figure 4.8 demonstrates an example of a Hough Transform based image segmentation. Base on the structure of the lines, we are able to efficiently segment the images into different sections. Each section often represents a paragraph or article within the image.

After segmentation, both logical and physical layouts of the document can then be understood. This can be used for reconstruction of the document as an electronic version. OCR or post-processing can also be applied more logically to each section.

The drawback for using this algorithm is that the existences of the lines within the images are compulsory before segmentation. As the logical core segment, the images are based on the layout of the lines. If required the algorithm can be keyed, an alternative parameterised feature.

4.4 Complex Newspaper Layout

Up until now, algorithms in *Ocropus* have been tested in this chapter on the historic newspapers that contain a rather straightforward layout (two columns with reasonable spaces in between). Nevertheless, newspaper layouts could frequently contain more than two columns, in addition to having much narrower spaces in between, which form a more complex newspaper layout structure.

Experiments have been performed, in order to understand how well the currently implemented algorithms can deal with such complex layout structures.

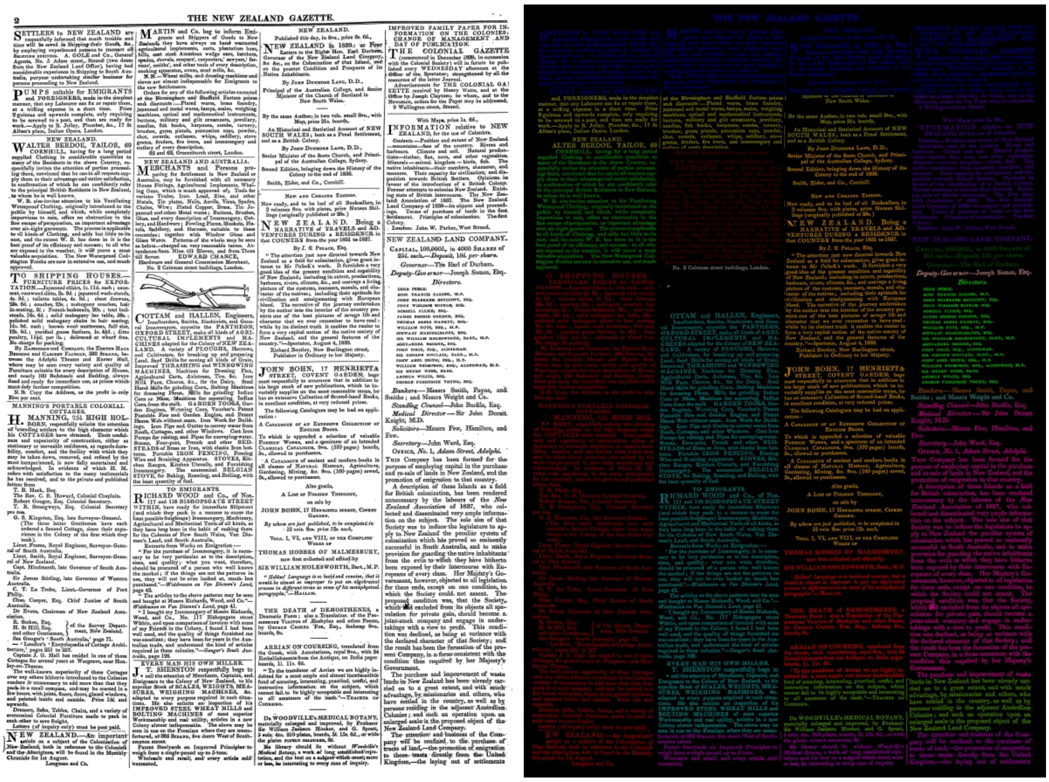


Figure 4.9: RAST segmentation with complex newspaper layout (column group, image dimension: 4000 x 5921, resolution: 96 DPI).

Figure 4.9 demonstrates an example for using RAST algorithm on a newspaper with more complex layout structures. This image is taken from the *Papers Past* website. Compared to previous examples the image not only has greater dimensions, it also consists of four columns. Each column contains several paragraphs and even has an illustrated drawing. Spaces between each column and paragraph also appear to be much narrower.

As it is shown in Figure 4.9, the RAST algorithm is able to segment the majority of paragraphs in each column. However, paragraphs have been mis-segmented on the top and bottom of the image. The algorithm has grouped paragraphs from different columns together, to make it into one paragraph, whereas they should belong to different columns as they have been separated by the vertical line. For this image RAST failed to group different paragraphs into a column, in addition 22% of image areas have been mis-segmented into different sections as indicated with different colours.

Furthermore as mentioned before, RAST removes those segments (pictures, lines), that it considered as non-text. This eliminates the problem for OCR non-text components in the image. In this case RAST also remove the majority of first letters in each paragraph, as the typography of this newspaper used extra large capital letters for the first character in each paragraph. Consequently, as the first character might not match with the following sentences on the same text line, it is likely to be removed by the algorithm.

We then performed the XY-Cut segmentation on the same image. Lines had been removed before hand by the line removal algorithm, because the current XY-Cut implementation is not yet capable of dealing with lines.

As it shown in Figure 4.10, only the title and paragraphs in the fourth columns have been segmented. The first three columns on the left hand side have been merged as one. It appears that the current XY-Cut algorithm is not yet mature enough for segmenting complex historical newspaper layouts as such. It appears that the XY-Cut algorithm is able to perform segmentation only on 26% of the image. Nevertheless, on the fourth column it shows the potential of what the XY-Cut can achieve.

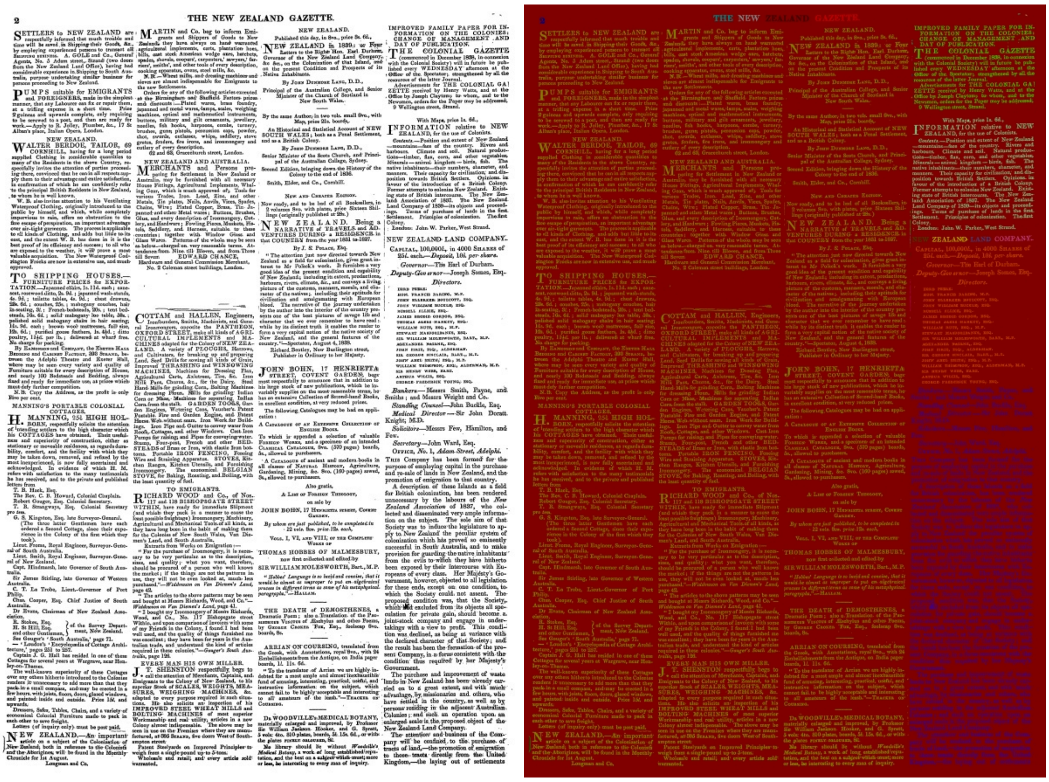


Figure 4.10: XY-Cut segmentation with complex newspaper layout (image dimension: 4000 x 5921, resolution: 96 DPI).

Segmentation is then performed on the same image using Hough Transform. As it is shown in Figure 4.11, the logic core has successively segmented the image into smaller sections, based on the structure of the lines. In this example, the algorithm has achieved 100% accuracy in segmenting the image into different sections relating to each article.

This demonstrates that the Hough Transform segmentation works exceptionally well on such an historical newspaper layout. In particular, for identifying sections of paragraphs or articles, as in the layout of the majority of historical newspapers horizontal lines are often used for the separation of the articles.

Furthermore, the position of columns within the newspaper can also be recognised. The algorithm can be modified for grouping each section into columns.

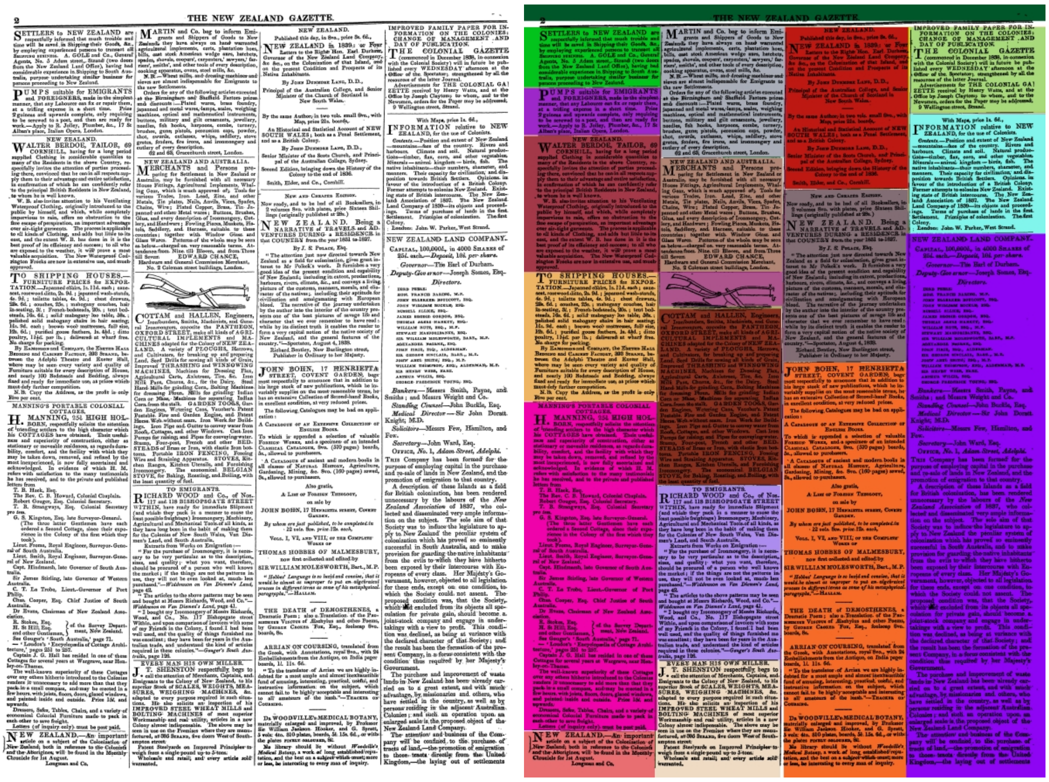


Figure 4.11: Hough Transform segmentation with complex newspaper layout (image dimension: 4000 x 5921, resolution: 96 DPI).

Experiments were performed on newspaper layouts with a higher degree of complexity, which consisted of seven columns. As is shown in Figure 4.12, the image is digitised from a full size newspaper. Its dimensions become a great factor during segmentation. Both RAST and XY-Cut algorithms failed to process the image due to its size. The memory appears to be insufficient for the algorithms to process images with such dimensions. Nevertheless, by using Hough Transform, columns can be extracted from the image.

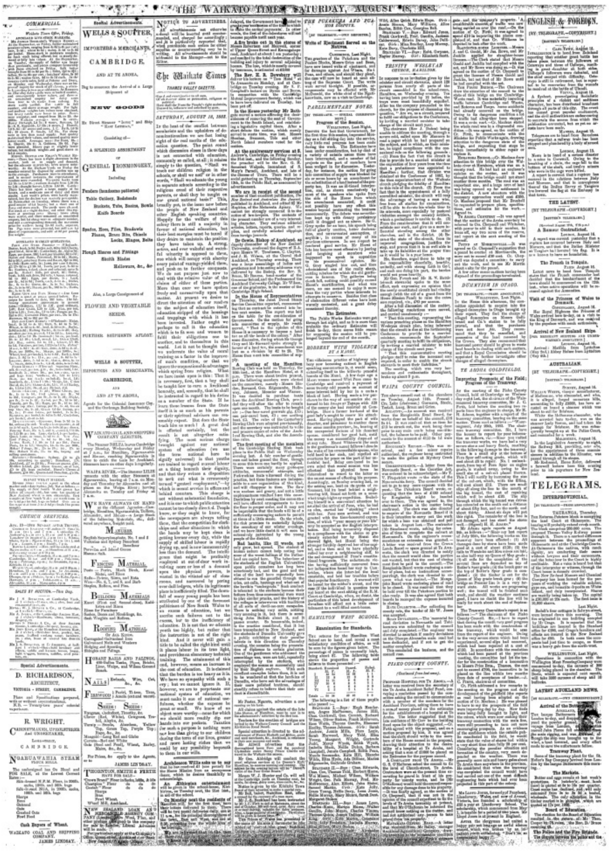


Figure 4.12: Complex newspaper layout with seven columns (image dimension: 7150 x 9921, resolution: 96 DPI).

In the context of an historic newspaper digital library, we cannot take only the moderate dimension into consideration, as the dimensions can vary, depending on the different newspapers published. Hence, segmentation algorithms should take different image dimensions into account during the implementation, to further enhance the utilisation of the memory usage during the document segmentation.

4.5 Different Scale of Images

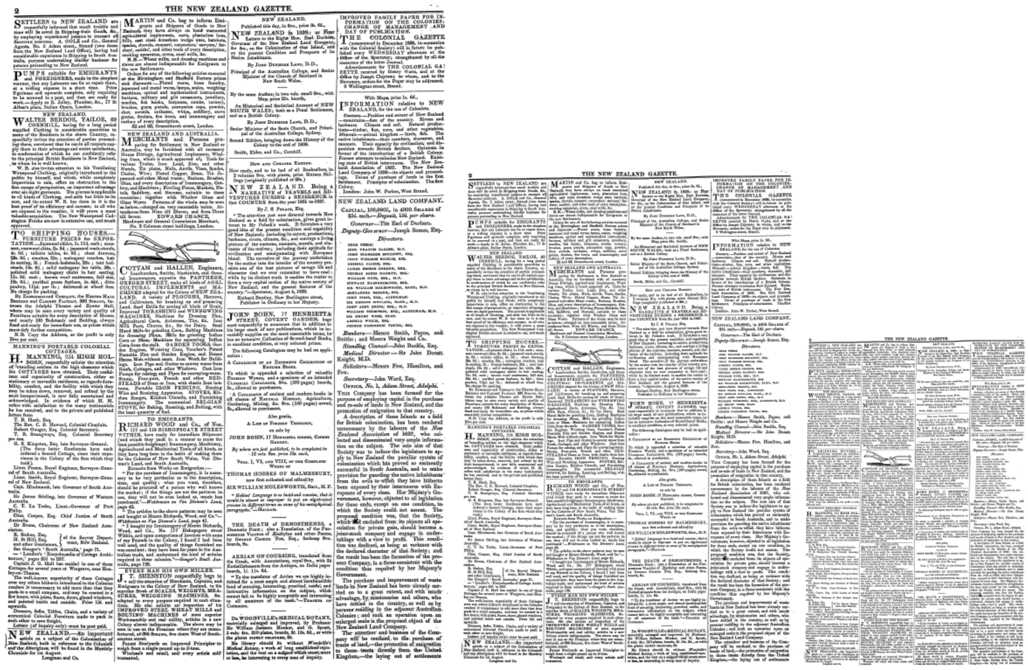


Figure 4.13: Different scale of the one image (original Image dimension: 4000 x 5921, resolution: 96 DPI).

As mentioned above, the Hough Transform algorithm is able to perform document segmentation, even with the image scaled down. Whereas the others struck difficulties in terms of resources needed. One way to alleviate the resources requirement is to work with scale reduced versions.

In order to evaluate the segmentation abilities between different algorithms over varying scales of one image, a set of scaled images from the image with initial dimensions of 4000 x 5921 has been created, as shown in Figure 4.13. A dimension of the next image is then reduced gradually, based on the original image. We then processed them with different algorithms for comparison.

For this experiment we consider the resulting areas on the image that failed to be segmented into a block of text or a paragraph by the algorithm as mis-segmentation.

Table 4.1 states the segmentation results for both Hough Transform and RAST algorithms. It shows Hough Transform is still capable of segmenting the image and extracting approximately 50% of sections even when the dimensions decrease to one tenth of its original size.

Dimension \ Algorithm	Hough Transform	RAST
4000 x 5921 (100%)	100% Segmentation	79% Segmentation
3000 x 4441 (75%)	100% Segmentation	78% Segmentation
2000 x 2961 (50%)	100% Segmentation	14% Segmentation
1000 x 1480 (25%)	100% Segmentation	7% Segmentation
400 x 592 (10%)	50% Segmentation	0% Segmentation

Table 4.1: Segmentation results over different scale of image.

As the Hough Transform depends on the line structure for performing segmentation, even with decreases in the dimensions of the image, line structure can still be recognised by the algorithm. Nevertheless, as the size decreases, pixels will start to be lost as well. In this example the segmentation accuracy for the Hough Transform starts to decrease when the dimensions are reduced to one tenth of the original image.

In comparison to the Hough Transform, RAST requires the image to be of higher quality. Because it takes text lines and white spaces into consideration, the accuracy therefore drops dramatically when the dimensions are reduced to half.

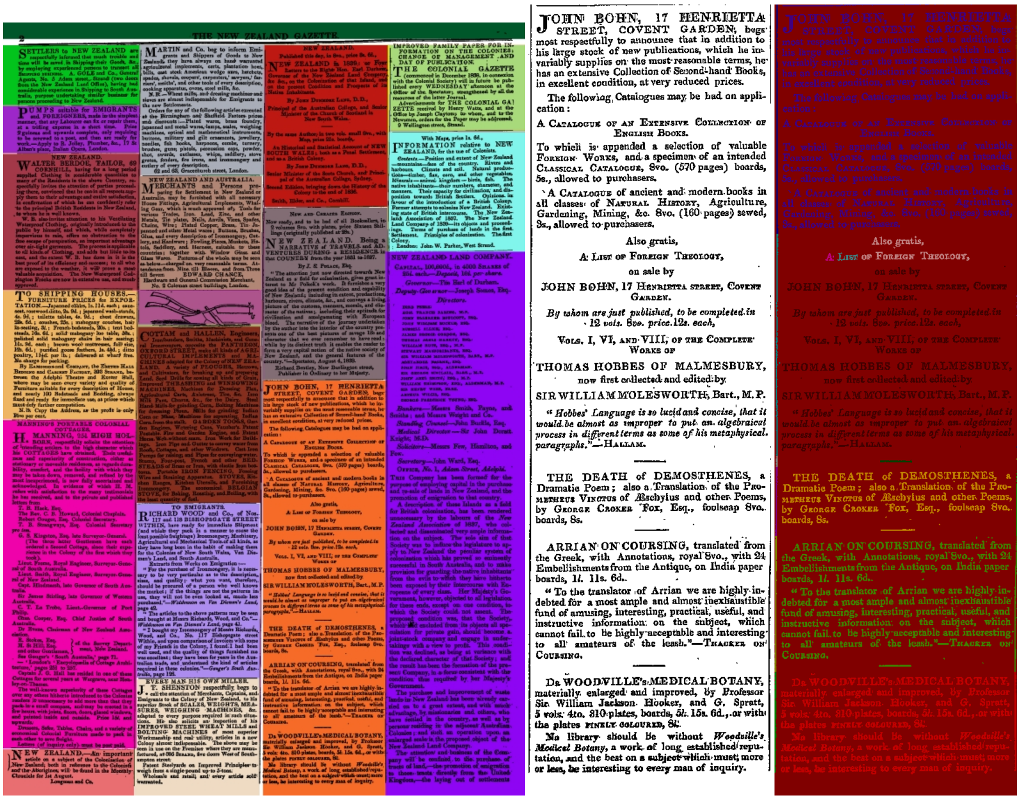
The character distortions make it quite difficult to be recognised by the human eye, especially when the dimensions are reduced to 25%.

In summary, for the example images, the Hough Transform has the higher degree of dimensional tolerance compared with RAST, this is further corroborated by testing across the corpus (see section 4.10). We can further use this characteristic in order to achieve more efficient segmentation for the Hough Transform.

4.6 Use of Combination

A series of experiments was conducted to identify and evaluating different characteristics for RAST, XY-Cut and the Hough Transform algorithms. Base on the segmentation results using the sample of historical newspapers, advantages and disadvantages for each algorithm have been established.

From the result of previous experiments, it demonstrates it can be challenging for established algorithms in *Ocropus* to achieve high accuracy segmentation on document layouts such as an historical newspaper. Hence, an experiment has been established by using the combination of algorithms for image segmentation.



JOHN BOHN, 17 HENRIETTA STREET, COVENT GARDEN, begs most respectfully to announce that in addition to his large stock of new publications, which he invariably supplies on the most reasonable terms, he has an extensive Collection of Second-hand Books, in excellent condition, at very reduced prices.

The following Catalogues may be had on application:

A Catalogue of an Extensive Collection of English Books.

To which is appended a selection of valuable Foreign Works, and a specimen of an intended Classical Catalogue, 8vo. (570 pages) boards, 5s., allowed to purchasers.

A Catalogue of ancient and modern books in all classes of Natural History, Agriculture, Gardening, Mining, &c. 8vo. (160 pages) sewed, 3s., allowed to purchasers.

Also gratis,

A List of Foreign Theology,

on sale by

JOHN BOHN, 17 Henrietta street, Covent Garden.

By whom are just published, to be completed in 12 vols. 8vo. price 12s. each,

Vols. I, VI, and VIII, of the Complete Works of

THOMAS HOBBES OF MALMESBURY,

now first collected and edited by

SIR WILLIAM MOLESWORTH, Bart., M.P.

"Hobbes' Language is so lucid and concise, that it would be almost as improper to put an algebraical process in different terms as some of his metaphysical paragraphs."—Hallam.

THE DEATH of DEMOSTHENES, a Dramatic Poem; also a Translation of the Prometheus Vinctus of Æschylus and other Poems, by George Croker Fox, Esq., foolscap 8vo. boards, 8s.

ARRIAN ON COURSING, translated from the Greek, with Annotations, royal 8vo., with 24 Embellishments from the Antique, on India paper boards, 1*l.* 11s. 6d.

"To the translator of Arrian we are highly indebted for a most ample and almost inexhaustible fund of amusing, interesting, practical, useful, and instructive information on the subject, which cannot fail to be highly acceptable and interesting to all amateurs of the leash."—Thacker on Coursing.

Dr WOODVILLE's MEDICAL BOTANY, materially enlarged and improved, by Professor Sir William Jackson Hooker, and G. Spratt, 5 vols. 4to. 310 plates, boards, 5*l.* 15s. 6d., or with the plates finely coloured, 8*l.*

No library should be without *Woodville's Medical Botany*, a work of long established reputation, and the best on a subject which must, more or less, be interesting to every man of inquiry.

Figure 4.14: Example of combination segmentation (original Image dimension: 3000 x 4441, resolution: 96 DPI).

As shown in Figure 4.14, the Hough Transform is first used for segmenting the newspapers into rectangular sections based on the line structure. Each section is then extracted from the image for further analysis by other algorithms.

The extracted section is then segmented by XY-Cut algorithm in Figure 4.14. By doing so the paragraphs in each section can then be further identified, as XY-Cut algorithm greatly enhances the strength in paragraph segmentation. As we can see, after extracting the orange section and putting it through the XY-Cut algorithm, the paragraphs are then segmented into different components. This provides us with a satisfactory layout of information of the section.

If we put the whole image instead of one section, into XY-Cut algorithm, it will involve a higher degree of difficulty for the algorithm to provide such satisfactory results.

Similarly, the RAST algorithm can also be applied to the section extracted. Due to its ability of identifying the text components from non-text, it will be additionally valuable for removing the non-text components within each section (drawing). This prevents the OCR engine from attempting to recognise the non-text components, which results in higher OCR accuracy.

Furthermore, segmentation can be carried out to a finer degree. By using other algorithms, position of each lines, words and even characters can be identified. From this experiment, benefits for using combinational algorithms on historical newspaper layouts can be understood, with both performance and segmentation accuracy being improved.

4.7 Evaluation Algorithm

Evaluation is an essential part for the development of an algorithm. A proper evaluation technique will allow us to examine the accuracy, performance and even the characteristics of an algorithm, enabling comparison to be made between different algorithms.

For this reason, a technique that provides a performance evaluation for algorithms against ground-truth data has been adopted (Phillips and Chhabra 1999) for the purpose of evaluating the algorithms used in our experiment. This technique has also been utilised in the First International Newspaper Segmentation Contest (Gatos et al. 2001) and many other similar projects.

The basis for this technique is counting the number of matches between the entities detected by the algorithm and the entities predefined in the ground-truth data. Multiple entities can be evaluated by using this technique, depending on the occurrence rate of each entity different weights can be assigned to the evaluation algorithm, which will affect the evaluation results. Only the same category of entity can be compared, i.e., the entity of line extracted can only be used to compare with, the entity of line defined in the ground-truth data.

For each image segmented, the segmentation result produced is weighted against the corresponding ground-truth data by the evaluation algorithm. As a result, it produces the counts of matched categories and status such as one-to-one matches, one-to-many matches, many-to-one matches, false-alarms and misses. False-alarms (erroneous detections) can be treated as entities in the segmentation results that do not match with any entity in the ground-truth data. Misses are entities in the ground-truth data that do not match with any entities in the segmentation results.

The Match Score Matrix is then produced from the evaluation results: scores are ranged from 0.0 to 1.0, 1.0 being a perfect match between segmentation results and the ground-truth data. Once the score is calculated we then create a Match-Count Table for counting the one-to-one matches. Thresholds have to be set in order to distinguish different matches.

This technique also provides different methods for comparing different entities, in order to calculate the match scores. For our experiment, we have adopted the method used for text area comparison, mentioned in the technique with slight modification for calculating the match score. By using this method we can compare the match for columns, paragraphs, lines of text and even for words.

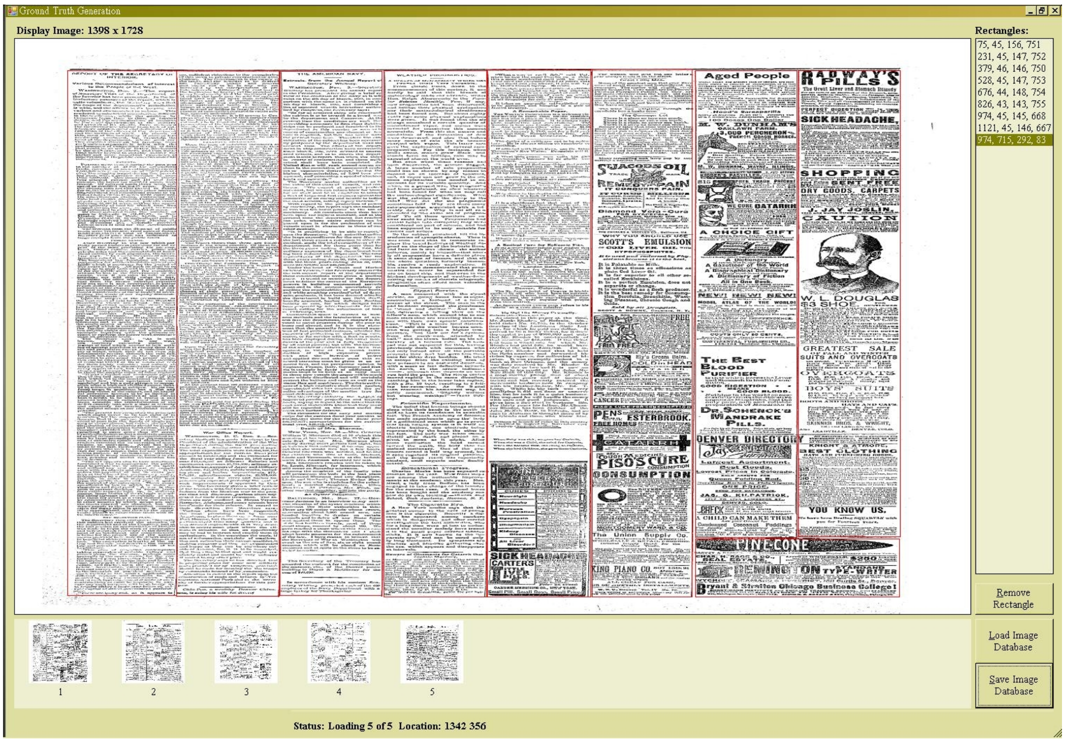
The method for text area comparison is calculated by first comparing both corners between the detected result and ground-truth data. If they are identical then we have a perfect match. Otherwise, we check whether they are overlapped, followed by calculating the intersection of D (detected result) and G (ground-truth). If there are no intersections we set the match score to zero, otherwise we calculate the area of D, G and $D \cap G$. The match score is then calculated by the following formula:

$$\mathrm{MatchScore} = \frac{\mathrm{area}(D \cap G)}{\max(\mathrm{area}(D),\mathrm{area}(G))}$$

For our experiment, the threshold has been set at 0.85 (suggested value in the article), and we consider two entities as a match with the match score being higher than the threshold. Nevertheless, in reality the threshold can be variable under different circumstances. The result detection rate for the algorithm is calculated by dividing the number of matches found by the total number of entities in ground-truth data. Percentage of missed detections and false-alarms are also computed in a similar manner.

For the following experiments, we put our focus on calculating one-to-one matches and the performance of the algorithm. The evaluation algorithm has added to *Ocropus*, along with minor modification of Hough Transform that allows the segmented sections to be grouped into columns. Due to the fact that column extraction has already been implemented in *Ocropus* for RAST, our evaluation will be based on extracting columns from the images in corpus.

Prior to evaluation, ground-truth data has to be created, and we have implemented simple software in C# for the generation of the ground-truth data as shown in Figure 4.15.



Ground Truth Generation

Display Image: 1398 x 1728

Rectangles:

75, 45, 156, 751
231, 45, 147, 752
379, 46, 146, 750
528, 45, 147, 753
676, 44, 148, 754
826, 43, 143, 755
974, 45, 145, 668
1121, 45, 146, 667
974, 715, 292, 83

Status: Loading 5 of 5 Location: 1342 356

Figure 4.15: Ground truth generation.

The software allows us to open a XML file that contains all five image names within each folder, we can than manually mark each column for each image. The marking column coordination is then outputted as a text file that can then be read into Lua scripts used by *Ocropus* to be used by the evaluation algorithm. The evaluation process was also been implemented in Lua scripts that allow each algorithm to traverse through all images within each collection. The result was then outputted as CSV (common separated value) files for later analysis.

4.8 Corpus Evaluation

Evaluation Result without Deskew										
Collection\Algorithm Accuracy		Hough Transform				RAST				
Collection Name	Resolution	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)	Fails
Australian Newspapers	3256x5119 300dpi	25.0	52.2	71.4	12.2	75	35	100	6.0	0
Austria Newspapers	3168x4066 96dpi	50.0	64.7	100.0	2.7	0	0	0	7.7	0
Brooklyn Daily Eagle	2544x3344 150dpi	16.7	63.3	100.0	3.3	N/A	N/A	N/A	N/A	5
Calgary Tribune	2129x3127 71dpi	12.5	7.9	12.5	2.4	0	0	0	8.1	0
California Digital Newspaper Collection	1558x2191 99dpi	44.4	61.4	85.7	2.3	0	0	0	25.6	1
Cape Vincent Eagle	2623x3821 150dpi	71.4	45.7	85.7	3.0	25	11.7	33.3	14.1	0
Casa Grande Dispatch	2331x3042 150dpi	4.5	20.0	120.0	4.7	4.5	1.5	4.5	9.9	0
Cherokee Phoenix	1185x1729 150dpi	60.0	80.0	100.0	1.6	0	0	0	14.9	0
Colorado	6301x7721 300dpi	11.1	41.7	75.0	12.6	N/A	N/A	N/A	N/A	5
Daily Enquirer	1695x2411 150dpi	50.0	61.5	85.7	2.8	0	0	0	11.8	1
Digital Library of the Caribbean	1138x1535 96dpi	20.0	50.0	100.0	1.5	0	0	0	8.5	0
Franklin Gazette	3727x4523 150dpi	36.4	57.1	77.8	5.0	N/A	N/A	N/A	N/A	5
Lethbridge Herald	4937x7054 300dpi	9.1	14.3	6.3	15.6	0	0	0	2.0	4
Logan Leader	3100x4067 100dpi	62.5	69.4	83.3	3.3	0	0	0	8.2	2
Newfoundland	1590x2211 150dpi	60.0	87.5	100.0	1.7	0	0	0	7.0	0
Niupepa-NZ	1143x2070 72dpi	50.0	70.0	100.0	0.6	0	0	0	0.6	0
PapersPast-NZ	4063x5354 96dpi	75.0	90.0	100.0	6.9	50	16.7	50	4.3	2
The British Colonist	1755x2488 150dpi	83.3	94.2	100.0	2.2	0	0	0	9.3	0
The Corrector	1513x2823 96dpi	33.3	64.5	83.3	3.6	0	0	0	17.6	0
The Long Islander	2137x2819 96dpi	20.0	52.0	60.0	2.6	0	0	0	13.5	0
Total Average			57.4		4.5		3.9		11.3	25

Table 4.2: Evaluation results without de-skew to corpus.

In this experiment we go over the evaluation algorithm through the corpus with Hough Transform and RAST segmentation algorithm without prior de-skew applied to the corpus. The trim edge algorithm has been applied to images with noises around the edges. Table 4.2 demonstrates the average match accuracy for five images in each collection, as well as the total amount of time for each collection. Through all collections, the Hough Transform encountered no difficulty while progressing through all images, while RAST has encountered problems in eight collections. In some collections a couple of images have failed to be processed by RAST, while others had failed all five images.

When RAST failed, an exception was thrown, which caused the evaluation algorithm to terminate prematurely. Our primary suspicion is that RAST has a problem dealing with the image over certain dimensions, that causes an insufficient memory error (2GB of RAM was used for these examples), resulting in the failing of the segmentation process. However, more experiments will be performed later on in this chapter to confirm such a fact.

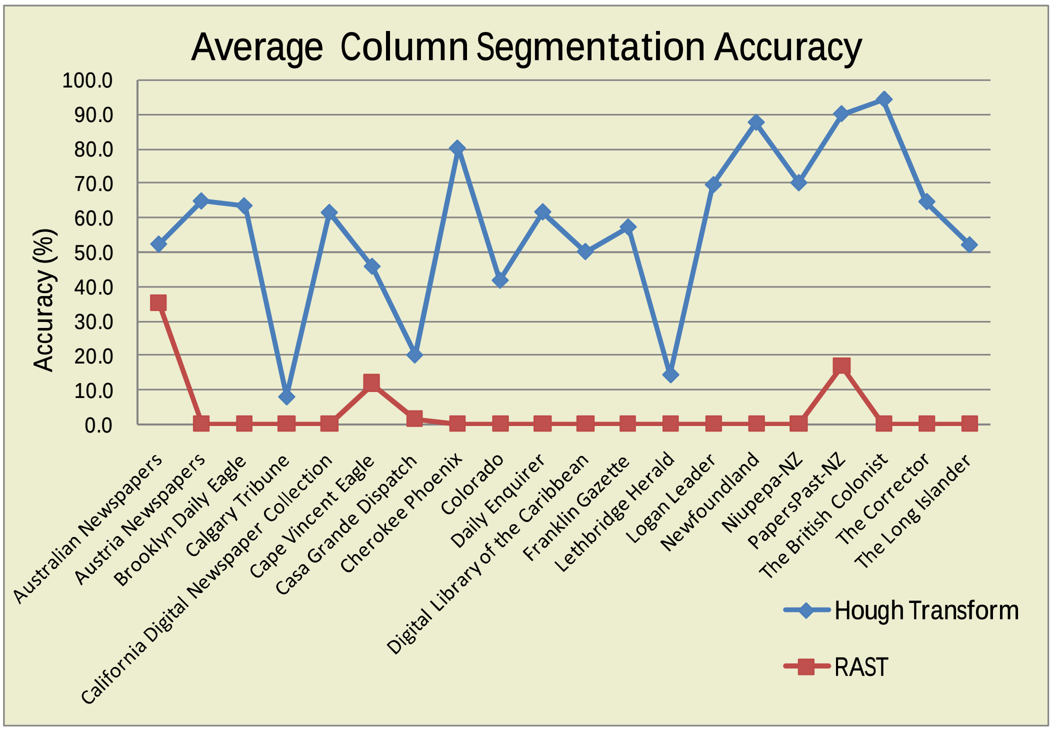


Figure 4.16: Average column segmentation accuracy.

Figure 4.16 shows the average column segmentation accuracy without de-skewing the corpus. For fifteen out of twenty collections Hough Transform achieved at least 50% of accuracy. Surprisingly RAST was only able to segment three collections into columns with the accuracy below 40%. Overall for this experiment Hough Transform has achieved average of 57.4% accuracy for column recognition, while RAST averages 3.9% (excluding failed images).

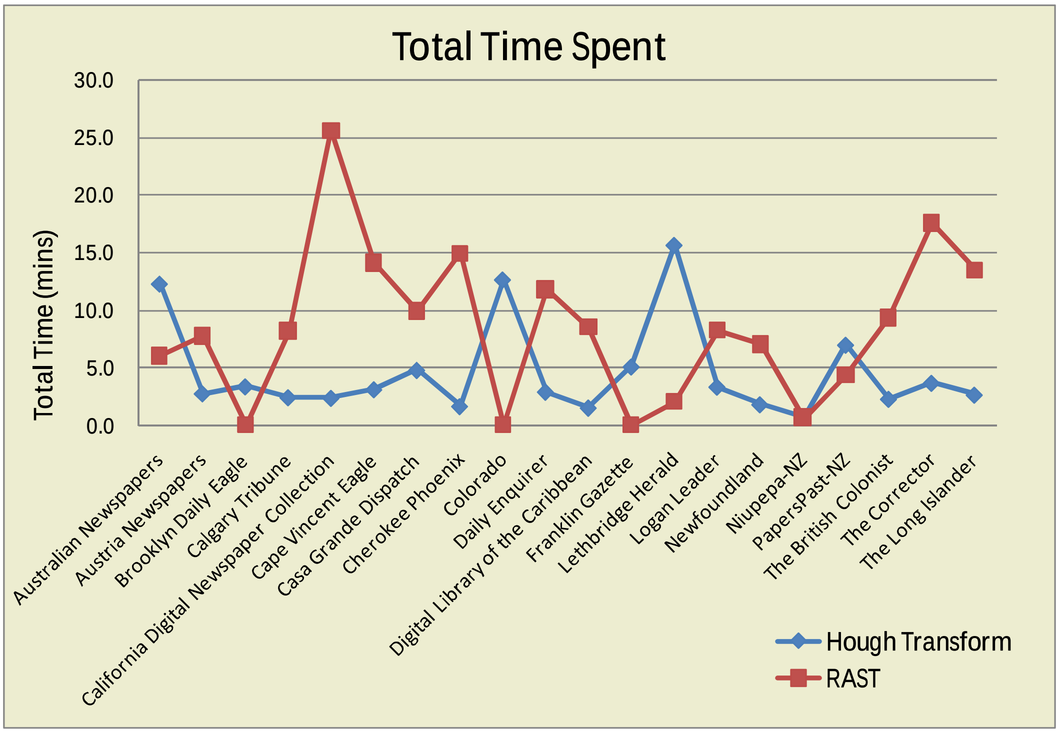


Figure 4.17: Total time spent for each collection.

In terms of performance, Figure 4.17 reveals the total time spent for both algorithms on each collection in the corpus. As is shown the Hough Transform is generally more efficient compare with RAST on most of the collections. In most cases we would expect an increase of time spent as the size of image increases, which means less time will be spent for an algorithm to process an image with a smaller size. However, this did not seem to be true for the California Digital Newspaper Collection. RAST spent a tremendous amount of time on this collection compared with other collections with similar image sizes. This pointed to the possibility that the performance of RAST might not be consistently in direct proportion to the dimension of the image. Further experiments were carried out to confirm this (see section 4.11). For this experiment The Hough Transform has the average of 4.5 minutes per collection (of 5 images) processed, and RAST on average, 11.3 minutes (excluding failed images).

4.9 Lower Scale Corpus Evaluation

In the previous experiment there were four collections that had at least four images which failed to be processed by RAST algorithm. Our speculation was that RAST crashed due to the fact that the image size in these collections was too large compared to images in other collections. In order to prove this we scaled down the image size in those four collections into lower resolution, and then performed the evaluation on both algorithms.

Lower Scale Evaluation without Deskew										
Collection\Algorithm Accuracy		Hough Transform				RAST				
Collection Name	Resolution	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)	Fails
Brooklyn Daily Eagle	1224x1582 72dpi	33.3	63.3	100.0	1.7	0.0	0.0	0.0	13.2	3
Colorado	1512x1853 72dpi	28.6	42.6	75.0	1.9	0.0	0.0	0.0	12.7	2
Franklin Gazette	1775x2160 72dpi	27.3	61.2	80.0	2.9	N/A	N/A	N/A	N/A	5
Lethbridge Herald	1196x1693 72dpi	14.3	7.8	37.5	2.1	0.0	0.0	0.0	8.7	0

Table 4.3: Lower scale evaluation without de-skew to corpus.

Table 4.3 illustrates the evaluation results gathered for both algorithms on those four collections. The result was unexpected. Originally we would have expected RAST to be able to process all the images without any difficulties, as the image size is now within a reasonable range compared to the images in other collections. RAST did improve by being able to process some of the images in some collections, but it still failed to process the majority of images.

From the results, it proves that the dimension of the image is not the only factor that causes RAST to fail. More importantly, layout structure of the newspaper and number of components within each image can also be an important factor that leads to failure of the algorithm.

Compared with RAST, the Hough Transform had no difficulty finishing processing the lower scale collection, with the addition of a slight increase in average accuracy on the first three collections, as well as decreases in the total time spent on all the collections.

4.10 De-skew Corpus Evaluation

In earlier experiments we noticed that the skew of an image can affect the segmentation result significantly. For this reason, an experiment was performed with de-skew of image applied, prior to evaluation. The De-skew process is automatically applied to images with skew angle greater or equal to 0.1 degree. Instead of using all twenty collections, for this experiment we restricted our attention to the collections that had no problem being processed by both Hough Transform and RAST algorithms.

Evaluation with Deskew														
Collection\Algorithm Accuracy			Hough Transform				RAST				Combination (Line Remove & RAST)			
Collection Name	Resolution	Deskew	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)	MIN (%)	Avg. (%)	MAX (%)	Total Time (mins)
Australian Newspapers	3256x5119 300dpi	Auto	50.0	56.5	71.4	12.2	14.3	27.9	50.0	9.8	14.3	27.9	50.0	9.3
Austria Newspapers	3168x4066 96dpi	Manual	50.0	70.6	100.0	3.3	50.0	5.9	50.0	5.3	50.0	5.9	50.0	5.4
Calgary Tribune	2129x3127 71dpi	Manual	37.5	7.5	37.5	3.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0	5.3
Cherokee Phoenix	1185x1729 150 dpi	Manual	60.0	72.0	100.0	1.8	0.0	0.0	0.0	6.7	0.0	0.0	0.0	6.5
Digital Library of the Caribbean	1138x1535 96dpi	Manual	20.0	41.7	60.0	1.2	0.0	0.0	0.0	7.5	0.0	0.0	0.0	7.4
Newfoundland	1590x2211 150dpi	Auto	60.0	91.7	100.0	1.7	0.0	0.0	0.0	6.5	0.0	0.0	0.0	6.6
Niupepa-NZ	1143x2070 72dpi	Auto	100.0	100.0	100.0	0.6	100.0	20.0	100.0	0.6	100.0	20.0	100.0	0.6
The British Colonist	1755x2488 150dpi	Manual	75.0	92.9	100.0	2.2	0.0	0.0	0.0	7.8	0.0	0.0	0.0	7.9
Cape Vincent Eagle	2623x3821 150dpi	Manual	71.4	45.7	80.0	3.4	25.0	18.3	66.7	15.6	50.0	23.3	66.7	16.7
Casa Grande Dispatch	2331x3042 150dpi	Manual	4.5	23.3	60.0	4.4	0.0	0.0	0.0	9.7	6.7	4.6	9.1	10.6
The Corrector	1513x2823 96dpi	Manual	42.9	41.9	66.7	3.2	0.0	0.0	0.0	16.5	0.0	0.0	0.0	16.9
The Long Islander	2137x2819 96dpi	Manual	60.0	60.0	60.0	2.0	0.0	0.0	0.0	11.8	0.0	0.0	0.0	11.7
Total Average				58.6		3.2		6.0		8.7		6.8		8.7

Table 4.4: Evaluation with de-skew to corpus.

Additionally, a third algorithm (a combination of the two) has been added into the evaluation. The combined algorithms first removed all the line components detected by Hough Transform from the image, followed by segmenting the image using RAST.

Table 4.4 shows the evaluation results from three algorithms along with their average accuracy and total time spent. For each collection, images have been de-skewed by using the RAST de-skew algorithms implemented in *Ocropus*. Nevertheless, result images cannot be guaranteed to be hundred percent correct. The current de-skew algorithm is not yet capable of correcting multiple angles of skew or distortion, that can occur quite frequently in historical newspaper. For this reason manual de-skew is often necessary. Where the de-skew algorithm did not perform satisfactorily, it was manually corrected, manually adjusted collections have been indicated in the table.

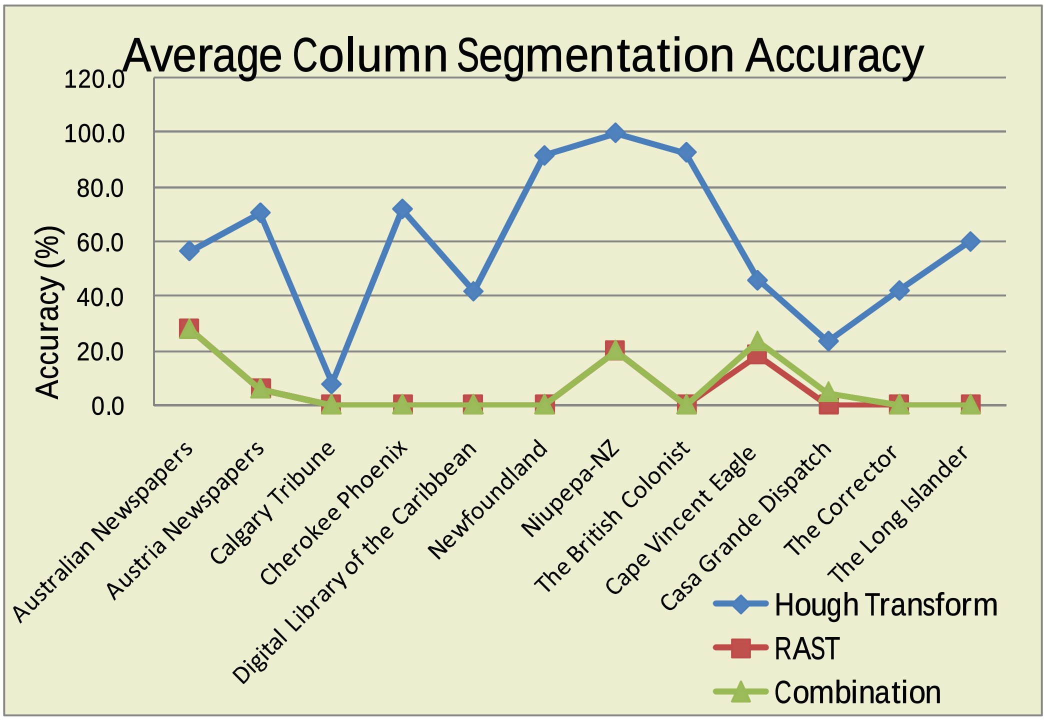


Figure 4.18: Average column segmentation accuracy with de-skew to corpus.

In this experiment the Hough Transform achieved an average of 58.6% over all twelve collections, compare with RAST with an average of 6% and Combination of 6.8%. Figure 4.18 illustrates the average column segmentation accuracy for all three algorithms. It seems that there are not major differences between the RAST and the combination algorithm. The results show that by using the combination algorithm accuracy for RAST increased, although the impact is insignificant at this stage.

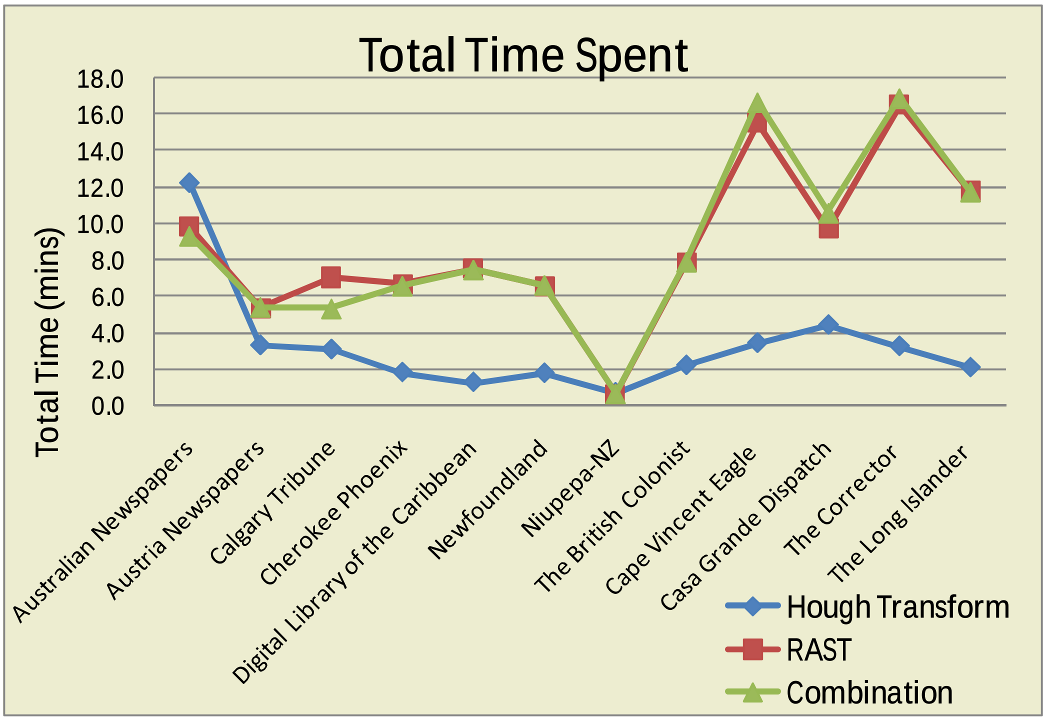


Figure 4.19: Total time spent with de-skew to corpus.

As for the performance for the de-skewed corpus, the Hough Transform reached an average of 3.2 minutes per collection, and RAST and Combination both with 8.7 minutes per collections. The total time spent for the combination algorithm is calculated after the removal of lines, as this is treated as image pre-processing. Overall for this experiment, the Hough Transform has achieved a better performance compared with the other two algorithms. By the comparison of

Combination and RAST, in most collections, performance has been slightly improved using the Combination algorithm.

It is apparent by removing the line components, we reduce the amount of elements needing to be processed by RAST. However, in two collections it seems there seems to be an exception. Nevertheless, in other collections it shows no significant difference between RAST and Combination algorithms.

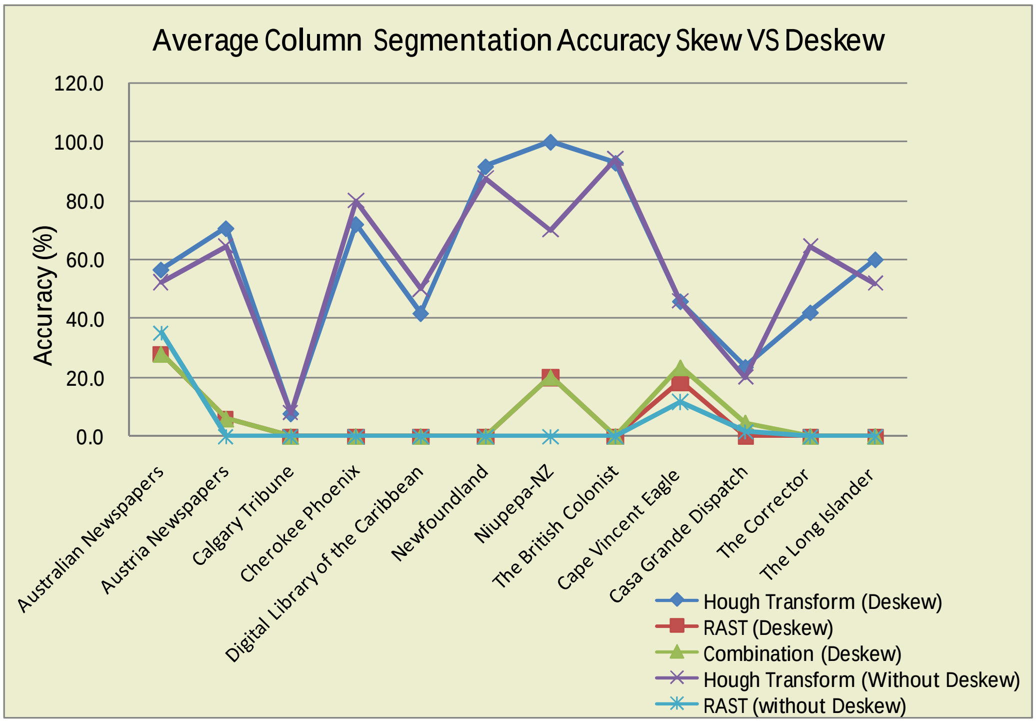


Figure 4.20: Average column segmentation accuracy skew VS de-skew.

To further examine the impact of skew on the segmentation accuracy, we have combined the segmentation results of the collection both before and after de-skew as shown in Figure 4.20. Generally speaking the accuracy is increased for all algorithms for most of the collections after de-skew. However, there are exceptions in some collections, because there are too much noises and multiple skew or distortions in the image. Once it has been de-skewed it will decrease the accuracy compared with the original angle.

This experiment proves that de-skew is an important pre-processing step to apply before segmentation of historical newspapers. Nevertheless, the current implementation of de-skew algorithms in *Ocropus* is yet to be perfected in dealing with multiple skew or distortions occurring in historic newspapers. To concentrate on the ideal performance of the segmentation algorithms in some cases we had to resort to manual correction.

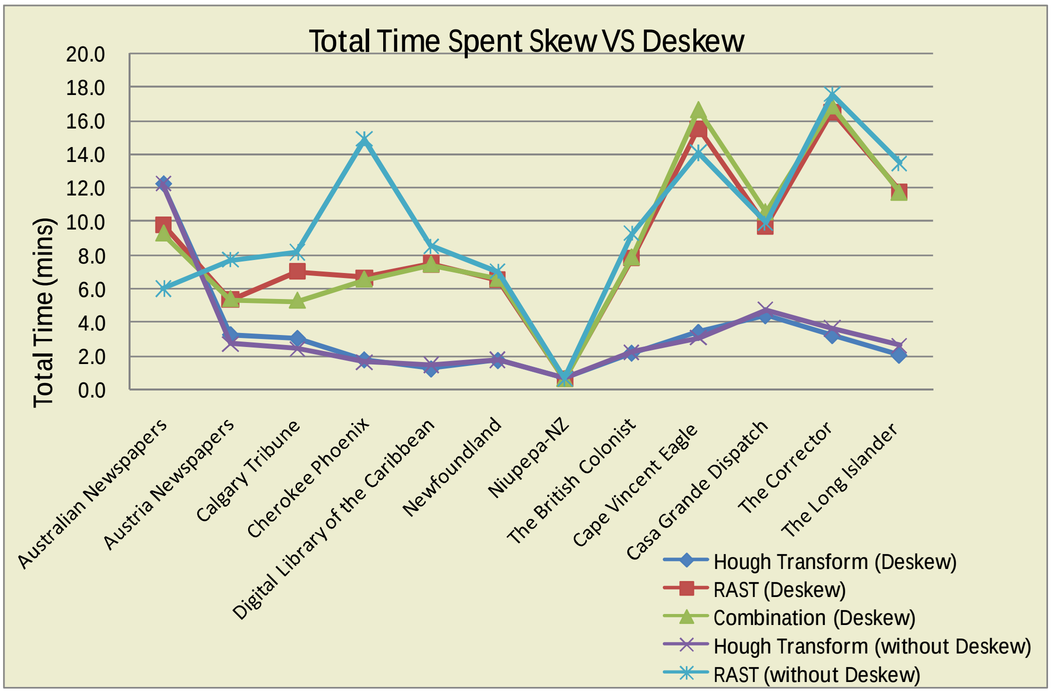


Figure 4.21: Total time spent skew VS de-skew.

Figure 4.21 illustrates the performance of the algorithms before and after de-skews. For the Hough Transform there are no major differences, as the performance of the algorithms is in proportion to the image size as well as the number of lines within the image. Therefore, even after de-skewing the number of lines within the image, it will retain approximately the same number of lines. For RAST the performance has increased significantly as the algorithm is based on text line and white space detection. Hence, after de-skew, the skew of the text line will be corrected, and white spaces between lines will be more evenly distributed, consequently improving the performance of RAST.

4.11 Scaling Corpus Evaluation

The process of building a collection of historic newspapers into a digital library can be extremely time consuming. It depends on the algorithms used, the dimension of the images, the number of the images within collections, and so on. Furthermore, for preservation purposes, images are often digitised in high resolution, this also causes an increased in processing time. However, in most cases segmentation can still be performed on lower resolution images with a certain accuracy trade-off, in order to achieve a better performance.

For this reason we have performed an experiment on one image from five collections in the corpus. These collections all have segmentation results returned by all algorithms from the previous experiment, except for the image in two of the collections that have no results from Hough Transform.

We have chosen the image with the highest segmentation accuracy within the collection. We start with the image dimension of 100% on both height and width and gradually reduce it by 25% at a time, until it reaches 25% of the original dimension. Evaluation is then performed on all different image scales, with the results shown in Table 4.5.

Different Scale Evaluation with Deskew (One Image from Each Collection)							
Collection\Algorithm Accuracy		Hough Transform		RAST		Combination (Line Remove & RAST)	
Collection Name	Resolution	Acuracy (%)	Total Time (mins)	Acuracy (%)	Total Time (mins)	Acuracy (%)	Total Time (mins)
Australian Newspapers	3256x5119 300dpi (100%)	50.0	2.3	25.0	1.3	25.0	1.3
Australian Newspapers	2442x3839 300dpi (75%)	25.0	1.4	0.0	1.3	0.0	1.3
Australian Newspapers	1628x2560 300dpi (50%)	25.0	1.8	0.0	1.7	0.0	1.8
Australian Newspapers	814x1280 300dpi (25%)	50.0	0.6	0.0	1.8	0.0	1.8
Austria Newspapers	3240x4122 96dpi (100%)	0.0	1.1	50.0	0.6	50.0	0.6
Austria Newspapers	2430x3092 96dpi (75%)	0.0	0.7	50.0	0.5	50.0	0.5
Austria Newspapers	1620x206196dpi (50%)	0.0	0.6	50.0	0.6	50.0	0.6
Austria Newspapers	810x1031 96dpi (25%)	0.0	0.2	0.0	0.7	0.0	0.7
Niupepa-NZ	1143x2070 72dpi (100%)	100.0	0.1	100.0	0.1	100.0	0.1
Niupepa-NZ	856x1532 72dpi (75%)	100.0	0.2	0.0	0.1	0.0	0.1
Niupepa-NZ	571x1022 72dpi (50%)	100.0	0.1	0.0	0.1	0.0	0.1
Niupepa-NZ	285x511 72dpi (25%)	100.0	0.0	0.0	0.2	0.0	0.2
Cape Vincent Eagle	1695x2809 150dpi (100%)	0.0	0.7	66.7	1.4	66.7	1.5
Cape Vincent Eagle	1271x2107 150dpi (75%)	50.0	0.2	16.7	1.2	0.0	1.3
Cape Vincent Eagle	848x1405 150dpi (50%)	0.0	0.2	0.0	0.9	0.0	1.0
Cape Vincent Eagle	424x702 150dpi (25%)	0.0	0.1	0.0	0.7	0.0	0.7
Casa Grande Dispatch	2337x3046 150dpi (100%)	4.5	0.9	0.0	1.3	9.1	1.5
Casa Grande Dispatch	1753x2285 150dpi (75%)	4.5	1.1	0.0	0.9	0.0	1.0
Casa Grande Dispatch	1169x1523 150dpi (50%)	4.5	0.5	0.0	0.7	0.0	0.7
Casa Grande Dispatch	584x762 150dpi (25%)	0.0	0.2	0.0	0.8	0.0	0.8

Table 4.5: Evaluation with image scaling to corpus.

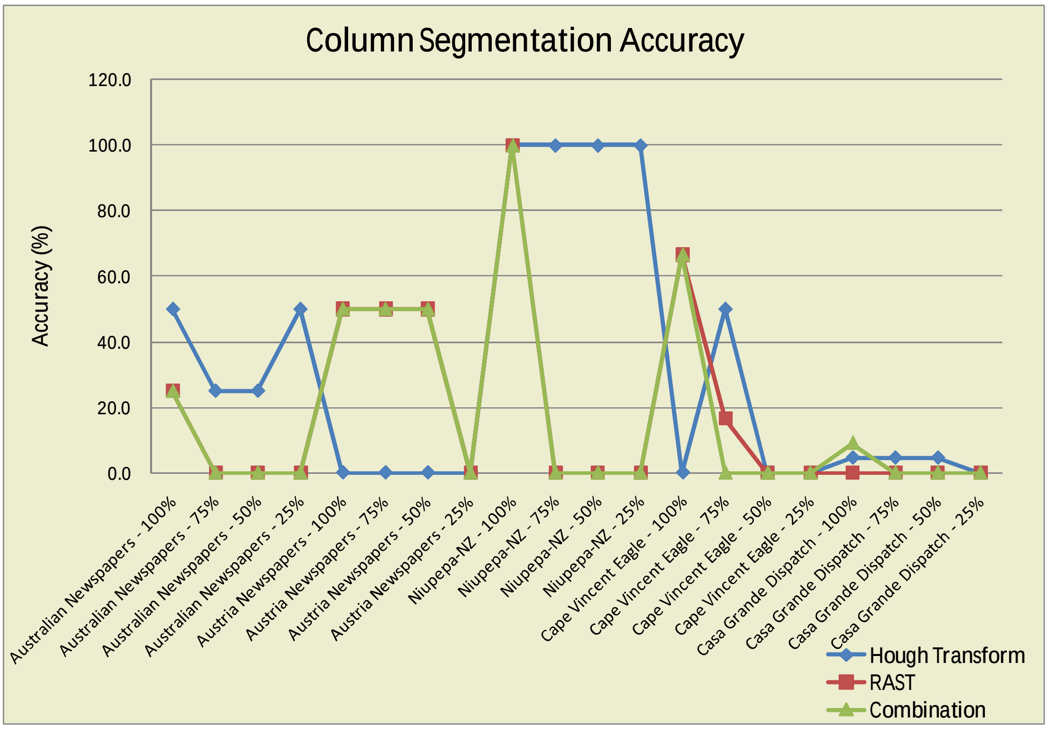


Figure 4.22: Segmentation accuracy on different scale.

Figure 4.22 demonstrates the segmentation accuracies over different image scaling on three algorithms. In two of the collections Hough Transform can still gather certain segmentation results even with the dimensions reduced to 25%. Most interestingly, for the collection of Cape Vincent Eagle, at 100%, Hough Transform was unable to gather any results. Yet when the dimensions were reduced to 75% some results had been found by Hough Transform. This was caused by the current implementation of the algorithm. Sometimes we have more accurate results when we scale down the images.

As for the RAST and Combination algorithms, three out of five of the collections results cannot be found once we reach 75% of image scaling. This proves that these algorithms, based on the image dimensions can be restrictive. Generally speaking the higher the resolutions the better the result.

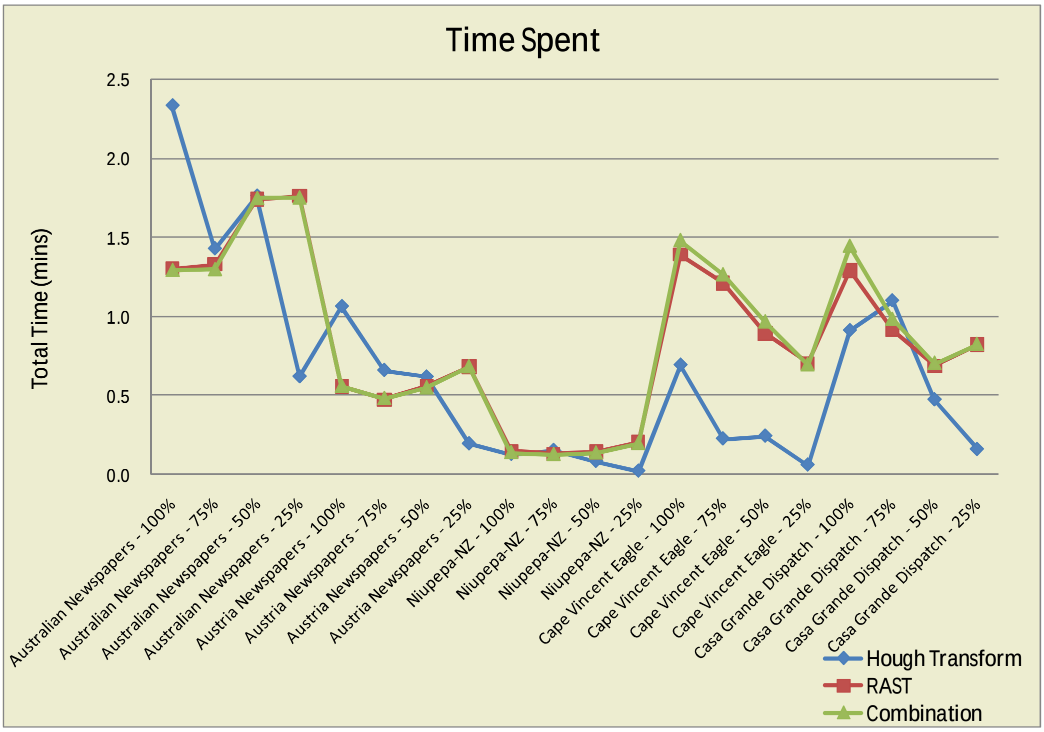


Figure 4.23: Time spent on different image scale.

Performance for processing different image scales for all algorithms can be found on Figure 4.23. For most of the collection, Hough Transform shows a trend of reducing the time spent as the image size reduces. However it is not always true for RAST based algorithms, on some images, the time spent increased as dimensions are reduced to 25%.

This experiment it shows the trade-off of the performance and accuracy of different algorithms on different scaling of images. For archiving historical newspapers using Hough Transform based algorithm, we could consider using 50% of image scaling as it significantly reduces the time spent, but we still maintain the same amount of accuracy from 75% or 100% of scaling. For using RAST based algorithms, it is we suggest, better to maintain the image with 100% scaling, as the accuracy starts to decrease as we reduce the dimension to 75%.

Nevertheless, for integrating segmentation algorithms into a DL system, more evaluation of the developed algorithms will be needed, in order to find a balance from performances and the resulting accuracies.

Chapter 5 - Summary and Conclusions

In Chapter 1 we introduced an example of an online digitised historic newspaper collection (Papers Past 2009) that was based using off-the-shelf commercial OCR software the OCR work was done separately to the formation of the Digital Library. In addition, Greenstone a comprehensive, open-source DL software system has been described. The software has the ability to provide a versatile platform for archiving vast amounts of historical treasures, such as historic newspapers into digital collections that can be more closely integrated with the document processing stage. Furthermore, it has a mechanism of plugins that can be built to order, to provide such processing during the process of ingesting such artefacts, rather than relying on an external process for document layout analysis and OCR.

In Chapter 2 several image-processing techniques and DLA algorithms have been introduced. Additionally numerous related projects were reviewed. While many of projects offer techniques or systems for archiving historic newspapers, nevertheless focus has often been aimed at only a few collections. Hence such a system can appear to be inadequate in providing authentic information from the newspapers for the integration of DL systems.

The development of a system or framework that can offer more authentic information, in favour of more generalised historical newspaper collections is indispensable, for the purpose of archiving a diversity of historic newspapers into digital collections.

Open-source software that can be used for the purpose of developing such a framework, has also been introduced, along with the reasons for choosing *Ocropus* as the implementation framework for our experiments.

In Chapter 3, implementation of the DLA algorithm, along with the different usages for line detection and removal, based on Hough Transform, was presented. Development of our own trim edge algorithm for better edge noises removal, has also been introduced. The corpus gathered from different online archives used for our experiments was also summarised in this chapter.

In Chapter 4, a selection of representative images with the various algorithms applied is given. Following this the analysis of applying the techniques to the assembled corpus is provided. Throughout the experiments different properties of DLA algorithms (RAST, XY-Cut and Hough Transform) were used and examined under a variety of circumstances (with/without skew, different scale and different newspaper layout). The adopted evaluation algorithms of Phillips and Chhabra (1999) that have been used by many other research projects in the past, were adopted for this. Additionally the process of generating ground-truth data for the corpus was explained.

Through this series of experiments we started to get an overview of how the current implemented DLA algorithms in *Ocropus* are yet to be proved adequate in providing segmentation to a document layout, such as the historic newspapers.

Compared with the major algorithms (RAST) provided by *Ocropus*, XY-Cut and Voronoi have not yet been fully developed. It has been proved through our experiments to be highly sensitive to skew and noises and segmentation accuracy is affected significantly. Nevertheless, skew can be frequently found in the digitised historic newspapers, caused during the digitisation process or by the deterioration of the newspaper.

Furthermore, current implementations of RAST have been proved to be memory consuming, compared with other algorithms tested, when apply to historic newspaper images. 25% of the images failed to be processed in the corpus due to insufficient memory. Due to its characteristics of using text line based detection and white spaces between each components, the image is often required to have a certain resolution to achieve high degree of accuracy. Therefore accuracy in segmentation will be greatly decreased with lower scale images.

Additionally, RAST proved to be time consuming when performing segmentation on the images in corpus. This is a drawback during the process of creating digital newspaper collections, as performance is a crucial factor, while the DL software system archives a vast amount of newspaper images.

Compared to RAST algorithm, our experiments demonstrates that the Hough Transform algorithm have higher column recognition accuracies throughout different historic newspaper collections. It achieved an average of 57.4% recognition rate compare to 3.9% by RAST through a total of twenty collections. Furthermore, it demonstrates a higher degree of skew tolerances compared to RAST. Segmentation results can still be retrieved within certain degrees of skew.

In terms of performance, Hough Transform takes an average of 4.5 minutes per collection (five images) compared with RAST, of 11.3 minutes making it almost 2.5 times more efficient.

The time spent overall for Hough Transform is more consistent, when compared with RAST. The performance is in proportion with the image dimension for Hough Transform, but that may not be the case for RAST, as it shows an inconsistent amount of time spent against the images in the corpus.

Furthermore Hough Transform demonstrates an advantageous point in segmenting different scales of images while maintaining a certain amount of accuracy. By using such a technique, performances of the DL software system can be greatly enhanced while archiving the historic newspapers.

In order to provide better support to the DL software system for historic newspaper collections, development of a framework is vital. Fortunately such framework (*Ocropus*) has been developed in recent years (Breuel, Kaiserlautern 2007). As an open-source system many additional efforts have been constantly contributed to *Ocropus* by the whole community. However, it is our view that the current difficulty for such framework, is lack of sufficient evaluation towards its developed algorithms.

For example the current de-skew algorithm is inadequate in dealing with multiple skew and distortion occurring on historic newspapers. A RAST algorithm produces a high amount of inaccuracies and is too inefficient in providing segmentation for an historic newspaper.

We believe enhanced development of algorithms can be achieved in order to provide better segmentation results to historic newspapers, in *Ocropus*. Our implementation of Hough Transform algorithms is one of the best examples, though it is yet far from perfect when processing other newspaper collections. Nevertheless further enhancements can be made in the future.

In conclusion, based on the experiments performed, we believe that by using *Ocropus* as the framework, algorithms can be developed, providing a greater support to the DL software system (Greenstone) for historic newspapers. Furthermore depending on the nature of the algorithms, supports for different layout of documents (scientific journals, hand written documents, and et cetera) can also be provided.

On the journey of the archiving of historic newspaper collections, this is not the end, but only the beginning.

References

A. Amin and S. Fisher (2000), “A Document Skew Detection Method Using the Hough Transform”, Pattern Analysis and Applications, Springer London, Vol. 3, No. 3, pp. 243-253.

B. T. Ávila and R. D. Lins (2005), “A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images”, Proceedings of the ACM Symposium on Document Engineering, pp. 118-126.

A. Bagdanov and J. Kanai (1997), “Projection Profile Based Skew Estimation Algorithm for JBIG Compressed Images”, ICDAR Proceedings of the 4th International Conference on Document Analysis and Recognition, pp. 401-406.

D. Bainbridge, S. Jones, S. McIntosh, M. Jones and I. H. Witten (2008), “Running Greenstone on an iPod”, International Conference on Digital Libraries, Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 333-336.

D. H. Ballard (1980), “Generalizing the Hough Transform to Detect Arbitrary Shapes”, Pattern Recognition, Vol. 13, No. 2, pp. 111-122.

J. E. Bresenham (1965), “Algorithm for Computer Control of a Digital Plotter”, IBM Systems Journal, Vol. 4, No. 1, pp. 25-30.

T. M. Breuel (1992), “Fast Recognition Using Adaptive Subdivision of Transformation Space”, IEEE Computer Vision and Pattern Recognition, pp. 445-451.

T. M. Breuel, DFKI and U. K. Kaiserslautern (2007), “The OCRopus Open Source OCR System”, Proceedings IS&T/SPIE 20th Annual Symposium, Vol. 6815.

R. Cattoni, T. Coianiz, S. Messelodi and C. M. Modena (1998), “Geometric Layout Analysis Techniques for Document Image Understanding: A Review”, ITC-IRST, Via Sommarive, I-38050 Povo, Trento, Italy.

G. S. Choudhury, T. DiLauro and R. Ferguson (2006), “Document Recognition for a Million Books”, D-Lib Magazine, Vol. 12, No. 3.

A. Downton, J. He and S. Lucas (2006), “User-Configurable OCR Enhancement for Online Natural History Archives”, International Journal on Document Analysis and Recognition, Vol. 9, No. 2, pp. 263-279.

R. O. Duda and P. E. Hart (1972), “Use of the Hough Transformation To Detect Lines and Curves in Pictures”, Communications of the ACM, Vol. 15, No. 1, pp. 11-15.

B. Gatos, S. L. Mantzairs, S. J. Perantonis and A. Tsigris (2000), “Automatic Page Analysis for the Creation of a Digital Library from Newspaper Archives”, International Journal on Digital Libraries, Vol. 3, No. 1, pp. 77-84.

B. Gatos, S. L. Mantzaris and A. Antonacopoulos (2001), “First International Newspaper Segmentation Contest”, Sixth International Conference on Document Analysis and Recognition, pp. 1190-1194.

Greenstone Plugins (2009), “http://wiki.greenstone.org/wiki/index.php/Plugins/”

T. Kanungo and R. B. Allen (1999), “Full-Text Access to Historical Newspapers”, LAMP-TR-033, CAR-TR-915, CS-TR-4014, MDA 9049-6C-1250.

S. Mori, C. Y. Suen and K. Yamamoto (1992), “Historical Review of OCR Research and Development”, Proceedings of the IEEE, Vol. 80, No. 7, pp. 1029-1058.

G. Nagy and S. Seth (1984), “Hierarchical Representation of Optically Scanned Documents”, 7th International Conference on Pattern Recognition, Vol. 1, pp. 347-349.

NATLIB NZ (2009), National Library of New Zealand Te Puna Matauranga O Aotearoa, “http://www.natlib.govt.nz/”

N. Otsu (1979), “A Threshold Selection Method from Gray-Level Histograms”, IEEE Transactions on Systems, Man and Cybernetics, Vol. 9, No. 1, pp. 62-66.

Papers Past Case Study (2008), Accessibility Issues, "http://www.dlconsulting.com/greenstone-services/marketing/GSpaperspastctp_final2.pdf"

Papers Past (2009), "http://paperspast.natlib.govt.nz/"

I. T. Phillip, S. Chen and R. M. Haralick (1993), "CD-ROM Document Database Standard", IEEE Proceedings of the Second International Conference on Document Analysis and Recognition, pp. 478-483.

I. T. Phillips and A. K. Chhabra (1999), "Empirical Performance Evaluation of Graphics Recognition Systems", IEEE Transactions Pattern Analysis and Machine Intelligence, Vol. 21, No. 9, pp. 849-870.

L. I. Rudin, S. Osher and E. Fatemi (1992), "Nonlinear Total Variation Based Noise Removal Algorithms", Physica D, Vol. 60, No. 1-4, pp. 259-268.

J. Sauvola and M. Pietikäinen (2000), "Adaptive Document Image Binarisation", Pattern Recognition, Vol. 33, No. 2, pp. 225-236.

J. Song and E. J. Delp (1992), "A Study of the Generalized Morphological Filter", Circuits, Systems, and Signal Processing, Vol. 11, No. 1, pp. 229-252.

A. L. Spitz (2003), "Correcting for Variable Skew in Document Image", International Journal on Document Analysis and Recognition, Vol. 6, No. 3, pp. 192-200.

The New Zealand Digital Library (NZDL) 2009, "http://www.nzdl.org/"

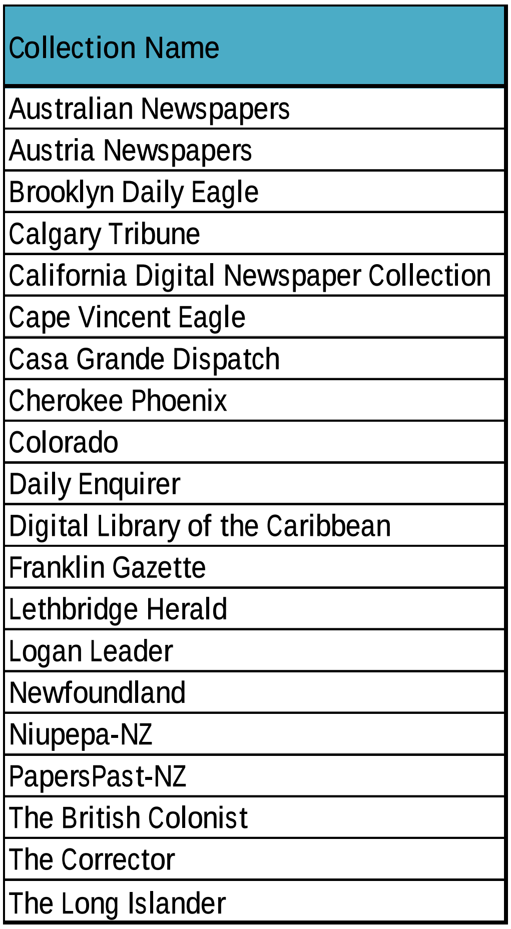
I. H. Witten, R. J. McNab, S. J. Boddie and D. Bainbridge (2000), "Greenstone: A Comprehensive Open-Source Digital Library Software System", International Conference on Digital Libraries, Proceedings of the 5th ACM conference on Digital libraries, pp. 113-121.

I. H. Witten and D. Bainbridge (2002), "How to Build a Digital Library 1st edition", Morgan Kaufmann, 340 Pine St, 6th floor San Francisco, CA 94104, USA.

Appendix

The attached DVD-ROM contains the corpus of our assembled historical newspapers, ground-truth data and experiment results. The corpus includes twenty historical newspaper collections, each collection containing five images in the collection folder.

Name of twenty collections as illustrated in Table A.1:



Collection Name
Australian Newspapers
Austria Newspapers
Brooklyn Daily Eagle
Calgary Tribune
California Digital Newspaper Collection
Cape Vincent Eagle
Casa Grande Dispatch
Cherokee Phoenix
Colorado
Daily Enquirer
Digital Library of the Caribbean
Franklin Gazette
Lethbridge Herald
Logan Leader
Newfoundland
Niupepa-NZ
PapersPast-NZ
The British Colonist
The Corrector
The Long Islander

Table A.1: Name of twenty historical newspaper collections.

In each folder five images is named from “1.PNG” to “5.PNG”, with the corresponded ground-truth data of columns stored in files named from “1.txt” to

“5.txt”. In each folder, we have saved the process of segmentation of different algorithms with its performance and evaluation results in HTML files. For example, “HoughTransform.html” contains the entire process of segmentation and evaluation of images within the folder of the Hough Transform algorithm. In other folders, evaluation results of all images using different algorithms has been saved in “algorithm name.csv” file, for example, “HoughTransform.csv” or “RAST.csv”.

Folders with different naming have been created for each section of the experiments, for section 4.9 Lower Scale Corpus Evaluation, folder name with the format of “collection name-Low” was created. For section 4.10 De-skew Corpus Evaluation with the folder name format of “collection name-Deskew”. Section 4.11 Scaling Corpus Evaluation with the folder name format of “collection name-Scale”. Each folder containing the particular evaluation results for the each experiment performed in each section.

File name, dimensions and resoultion for each image within a collection has been shown in Table A.2 with the collection name listed in alphabatical orders.

Image Dimension and Resolution		
Australian Newspapers		
File Name	Dimensions	Resolution
1.PNG	3256 x 5119	300dpi
2.PNG	3256 x 5119	300dpi
3.PNG	3256 x 5119	300dpi
4.PNG	3256 x 5119	300dpi
5.PNG	4423 x 5569	300dpi
Austria Newspapers		
File Name	Dimensions	Resolution
1.PNG	1280 x 2004	96 dpi
2.PNG	3168 x 4066	300dpi
3.PNG	1280 x 2013	96 dpi
4.PNG	1024 x 1565	96 dpi
5.PNG	3168 x 4066	300dpi
Brooklyn Daily Eagle		
File Name	Dimensions	Resolution
1.PNG	2551 x 3295	150 dpi
2.PNG	2554 x 3330	150 dpi
3.PNG	2539 x 3312	150 dpi
4.PNG	2544 x 3344	150 dpi
5.PNG	2551 x 3294	150 dpi
Calgary Tribune		
File Name	Dimensions	Resolution
1.PNG	2787 x 3127	71 dpi
2.PNG	2129 x 3127	71 dpi
3.PNG	2129 x 3127	71 dpi
4.PNG	2129 x 3127	71 dpi
5.PNG	2129 x 3127	71 dpi
California Digital Newspaper Collection		
File Name	Dimensions	Resolution
1.PNG	1055 x 1457	96 dpi
2.PNG	1558 x 2191	96 dpi
3.PNG	1441 x 1929	99 dpi
4.PNG	1441 x 1929	99 dpi
5.PNG	1441 x 1929	99 dpi

Table A.2: All Image dimension and resoultion in each collection.

Cape Vincent Eagle		
File Name	Dimensions	Resolution
1.PNG	1685 x 2803	150 dpi
2.PNG	1647 x 2738	150 dpi
3.PNG	2461 x 3237	150 dpi
4.PNG	2510 x 3213	150 dpi
5.PNG	2299 x 3052	150 dpi
Casa Grande Dispatch		
File Name	Dimensions	Resolution
1.PNG	2331 x 3037	150 dpi
2.PNG	2331 x 3042	150 dpi
3.PNG	2331 x 3042	150 dpi
4.PNG	2331 x 3042	150 dpi
5.PNG	2331 x 3042	150 dpi
Cherokee Phoenix		
File Name	Dimensions	Resolution
1.PNG	1185 x 1729	150 dpi
2.PNG	1275 x 1650	150 dpi
3.PNG	1165 x 1750	150 dpi
4.PNG	1169 x 1750	150 dpi
5.PNG	1177 x 1754	150 dpi
Colorado		
File Name	Dimensions	Resolution
1.PNG	5825 x 7199	300 dpi
2.PNG	5787 x 7369	300 dpi
3.PNG	6126 x 7611	300 dpi
4.PNG	5751 x 7141	300 dpi
5.PNG	6301 x 7721	300 dpi
Daily Enquirer		
File Name	Dimensions	Resolution
1.PNG	1675 x 2377	150 dpi
2.PNG	1634 x 2389	150 dpi
3.PNG	1675 x 2389	150 dpi
4.PNG	1661 x 2362	150 dpi
5.PNG	1695 x 2411	150 dpi

Table A.2: All Image dimension and resoultion in each collection.

<table>
<tr><th colspan="3">Digital Library of the Caribbean</th></tr>
<tr><th>File Name</th><th>Dimensions</th><th>Resolution</th></tr>
<tr><td>1.PNG</td><td>1055 x 1535</td><td>96 dpi</td></tr>
<tr><td>2.PNG</td><td>1093 x 1535</td><td>96 dpi</td></tr>
<tr><td>3.PNG</td><td>910 x 1500</td><td>96 dpi</td></tr>
<tr><td>4.PNG</td><td>917 x 1500</td><td>151 dpi</td></tr>
<tr><td>5.PNG</td><td>1138 x 1535</td><td>96 dpi</td></tr>
<tr><th colspan="3">Franklin Gazette</th></tr>
<tr><th>File Name</th><th>Dimensions</th><th>Resolution</th></tr>
<tr><td>1.PNG</td><td>3727 x 4523</td><td>150 dpi</td></tr>
<tr><td>2.PNG</td><td>3265 x 4031</td><td>150 dpi</td></tr>
<tr><td>3.PNG</td><td>3698 x 4500</td><td>150 dpi</td></tr>
<tr><td>4.PNG</td><td>3627 x 4442</td><td>150 dpi</td></tr>
<tr><td>5.PNG</td><td>3665 x 4473</td><td>150 dpi</td></tr>
<tr><th colspan="3">Lethbridge Herald</th></tr>
<tr><th>File Name</th><th>Dimensions</th><th>Resolution</th></tr>
<tr><td>1.PNG</td><td>4888 x 7004</td><td>300 dpi</td></tr>
<tr><td>2.PNG</td><td>4767 x 7004</td><td>300 dpi</td></tr>
<tr><td>3.PNG</td><td>4521 x 6617</td><td>300 dpi</td></tr>
<tr><td>4.PNG</td><td>4937 x 7054</td><td>300 dpi</td></tr>
<tr><td>5.PNG</td><td>4983 x 7054</td><td>300 dpi</td></tr>
<tr><th colspan="3">Logan Leader</th></tr>
<tr><th>File Name</th><th>Dimensions</th><th>Resolution</th></tr>
<tr><td>1.PNG</td><td>3100 x 4067</td><td>99 dpi</td></tr>
<tr><td>2.PNG</td><td>3013 x 4128</td><td>99 dpi</td></tr>
<tr><td>3.PNG</td><td>3040 x 4056</td><td>99 dpi</td></tr>
<tr><td>4.PNG</td><td>3067 x 4063</td><td>99 dpi</td></tr>
<tr><td>5.PNG</td><td>3022 x 4142</td><td>99 dpi</td></tr>
<tr><th colspan="3">Newfoundland</th></tr>
<tr><th>File Name</th><th>Dimensions</th><th>Resolution</th></tr>
<tr><td>1.PNG</td><td>1590 x 2211</td><td>150 dpi</td></tr>
<tr><td>2.PNG</td><td>1550 x 2211</td><td>150 dpi</td></tr>
<tr><td>3.PNG</td><td>1570 x 2258</td><td>150 dpi</td></tr>
<tr><td>4.PNG</td><td>1561 x 2231</td><td>150 dpi</td></tr>
<tr><td>5.PNG</td><td>1467 x 2128</td><td>150 dpi</td></tr>
</table>

Table A.2: All Image dimension and resoultion in each collection.

Niupepa-NZ		
File Name	Dimensions	Resolution
1.PNG	1141 x 2043	96 dpi
2.PNG	1143 x 2070	96 dpi
3.PNG	739 x 794	96 dpi
4.PNG	770 x 1266	96 dpi
5.PNG	792 x 896	96 dpi

PapersPast-NZ		
File Name	Dimensions	Resolution
1.PNG	4063 x 5354	96 dpi
2.PNG	4063 x 5354	96 dpi
3.PNG	4063 x 5354	96 dpi
4.PNG	4000 x 5921	96 dpi
5.PNG	7150 x 9921	96 dpi

The British Colonist		
File Name	Dimensions	Resolution
1.PNG	1122 x 1586	150 dpi
2.PNG	1124 x 1595	150 dpi
3.PNG	1769 x 2516	150 dpi
4.PNG	1769 x 2493	150 dpi
5.PNG	1755 x 2488	150 dpi

The Corrector		
File Name	Dimensions	Resolution
1.PNG	1554 x 2779	96 dpi
2.PNG	1583 x 2637	96 dpi
3.PNG	1580 x 2772	96 dpi
4.PNG	1510 x 2852	96 dpi
5.PNG	1513 x 2823	96 dpi

The Long Islander		
File Name	Dimensions	Resolution
1.PNG	2137 x 2819	96 dpi
2.PNG	2137 x 2819	96 dpi
3.PNG	2137 x 2819	96 dpi
4.PNG	2137 x 2819	96 dpi
5.PNG	1993 x 2721	96 dpi

Table A.2: All Image dimension and resoultion in each collection.

Corpus DVD-ROM