

A Document Page Layout Analysis Technique: Enhanced Scientific Document Understanding Through the Combination of Salient Mathematical Features into a Support Vector Machine Trained for Equation Detection

in Partial Fulfillment of the Requirements for the Degree
M.S. in Computer Engineering

Presented to the Faculty of ECE
of Virginia Tech by

Jake Bruce

in July 2013

Supervisor: Prof. Lynn Abbott, Ph.D.
Co-Supervisors: Prof. Jason Xuan, Ph.D. and Prof. Jules White, Ph.D.

Introductory remarks

Version <number>, published in <month> <year>

© <year> <first name> <last name> (►<e-mail>)

Licence. This work is licensed under the Creative Commons Attribution - No Derivative Works 2.5 License. To view a copy of this license, visit ►<http://creativecommons.org/licenses/by-nd/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA. Contact the author to request other uses if necessary.

Trademarks and service marks. All trademarks, service marks, logos and company names mentioned in this work are property of their respective owner. They are protected under trademark law and unfair competition law.

The importance of the glossary. It is strongly recommended to read the glossary in full before starting with the first chapter.

Hints for screen use. This work is optimized for both screen and paper use. It is recommended to use the digital version where applicable. It is a file in Portable Document Format (PDF) with hyperlinks for convenient navigation. All hyperlinks are marked with link flags (►). Hyperlinks in diagrams might be marked with colored borders instead.

Navigation aid for bibliographic references. Bibliographic references to works which are publicly available as PDF files mention the logical page number and an offset (if non-zero) to calculate the physical page number. For example, to look up [Example :a01, p. 100₋₈₀] jump to physical page 20 in your PDF viewer.

Declaration

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

Up to now, this thesis was not published or presented to another examinations office in the same or similar shape.

<place>, <date>

place and date

signature (<first name> <last name>)

Abstract

Abstract

<first paragraph title>. <The abstract of the diploma thesis is meta information resp. "management information". Therefore it should cover at most two pages, sum up the thesis' essentials and contain the idea behind the thesis.>

Acknowledgments

Acknowledgments

<Here is one page to thank everybody who helped and supported the author to write this thesis.>

Table of contents

Table of contents

	Abstract	v
	Acknowledgments	vii
	Table of contents	ix
1	Introduction	1
1.1	Enhancing Information Accessibility through Document Analysis and Recognition.....	1
1.2	Introduction to OCR and Document Analysis: A Brief History.....	5
1.3	Google's Open Source Initiative.....	8
1.4	Contributions of this Thesis.....	10
1.5	Organization of Thesis.....	11
2	Literature Review	13
2.1	Text styles.....	13
2.2	Text styles.....	13
1.1	Special text styles for patterns.....	15
1.1	Lists.....	16
1.1	Linking and referencing.....	17
1.1	Citing and bibliography.....	18
1	Object demonstration	21
A	Glossary of terms and abbreviations	I
B	Source listings	III
B.1	http_post().....	III
	Index of glossary items	V
	Index of objects	VII
	Bibliography	IX
	Colophon	XI
	Attached electronic data	XIII

1 Introduction

Basically, our goal is to organize the world's information and to make it universally accessible and useful.

Larry Page - Co-founder of Google

1.1 Enhancing Information Accessibility through Document Analysis and Recognition

Never, since the invention of the printing press, has society seen such a radical change in its means of information distribution. Armed with powerful search engines roaming the vast expanse of the World Wide Web, nearly everyone in the world has, at their very fingertips, access to archives full of information. This enhanced information accessibility is having profound implications for society and could lead to a fruitful age of enlightenment.

The global effects of high speed Internet access are seen daily as hundreds of millions browse for information/multimedia, look up map directions, interact through email/social networks/video games, shop remotely, video chat, etc. Corporations like Google, Microsoft, Facebook, eBay, and Amazon continue building and extending the capacity of their server farms as the growth of user demand shows no signs of slowing down. By mid-2012, it was reported that nearly an eighth of the world's population was on the popular social networking site, Facebook [1]. As such figures continue to grow, studies are showing that technology is even affecting the manner in which we think and behave at the most fundamental levels. Whether or not the long-term effects of this relatively nascent medium of interaction prove to be largely positive or negative remains to be seen. One remaining certainty, however, is that continuing innovation is, for better or for worse, altering the manner in which we live out our daily lives.

It was Benjamin Franklin who once said that “genius without education is like silver in the mine.” One would be hard-pressed in arguing that, throughout history, all people have been able to realize their full potential to succeed and make a difference in the world. If that were true, many would argue that our knowledge would, by now, have long since surpassed its current state. In fact it was just under five hundred years ago, that Europeans were finally emerging from an age of intellectual darkness which had lasted for roughly a millennium. If we look back to the spread of knowledge throughout written history, starting from the earliest true writing systems developed

1 Introduction

in ancient Egypt/Mesopotamia circa 3000 BC to the origins of philosophy, math, science, and theater in ancient Greece, all the way to the birth of the “modern era” which culminated itself in the scientific revolution of the sixteenth century AD, we notice a general trend of small bursts of knowledge spreading repeatedly, each time with greater strength than before, each one improving upon its predecessor. Sir Isaac Newton exemplified this trait of humanity with his statement that “if I have seen further, it is by standing on the shoulders of giants.”

Although much of what defines us from a cultural perspective may indeed be passed from generation to generation through word of mouth, our tremendous advancements in math, science, art, and literature since the dawn of the modern era can be largely attributed to Johannes Gutenberg's invention of the printing press, which made mass distribution of books possible in Late-Medieval Europe. Prior to this key event in history, the stage was set in Europe for an age of scientific inquiry and revelation when the religious leader, Thomas Aquinas, embraced the separation between the purely theological and purely scientific schools of thought. Also of vital importance was the translation and recurrence of ancient Greek writings which had been studied and further developed by Arabic scholars. The first universities built in Medieval Europe were initially centered around classical Greek and religious studies and helped to lead Europe out of its age of darkness. This collaborative environment of scholastic endeavor helped set the framework for an age of enlightenment which would move humanity a step forward. Archaic ideas such as bloodletting were soon supplanted by discoveries leading to modern medicine and the commonly held geocentric model of our earth was replaced by a heliocentric one. Major breakthroughs were made in every field to foster the spread of knowledge which took society to where it is today. Without this ideal of scientific thinking combined with the means to distribute information, society would have never seen such tremendous improvements.

Moving forward to the present day, society has recently made technological breakthroughs which make the world's knowledge and information more accessible than ever before. In fact, many have suggested that the widely used search engine, Google, will go down in history as rivaling in importance with Gutenberg's printing press. It was only about a decade and a half ago that two Stanford Ph.D. students decided that they would like to take a shot at downloading and categorizing the entire internet. These two graduate students are of course the founders of Google [2], a now successful multinational corporation which, during the late nineties, left its search

1 Introduction

engine competitors far behind. Google is unique in that its employees facilitate a diverse range of interesting projects ranging from cataloging the human genome, building autonomous vehicles, developing smart homes of the future, to developing augmented reality eye glasses, among many others. It is, however, in Google's core mission of finding ways to make the world's information "more universally accessible and useful," that the company has had its greatest impact on the world at large. It was in keeping with this mission that, in 2005, in collaboration with HP Labs and the Information Science and Research Institute at UNLV, Google revived and open sourced an optical character recognition engine that had been developed as a Ph.D. project for HP Labs between 1985 to 1995. Although optical character recognition (OCR), the autonomous conversion of printed documents into digital formats, is a very mature area of research [3], development in this area continues in order to increase recognition support for the broad spectrum of languages, formats, and subject matter of printed documents. HP's OCR engine, named "Tesseract," had proven itself as one of the industry's leading engines during UNLV's Fourth Annual Test of OCR Accuracy [4]. Eventually, however, HP subsequently went out of the OCR business, leaving the software to basically collect dust for about a decade.

Meanwhile, by around 2004, Google had begun its Google Books Initiative [5], a large-scale library digitization project. This initiative began with the lofty goal of digitizing all of the world's printed documents such that they may be indexed and searched online. By around 2005, Google hired Ray Smith, the former lead developer of Tesseract, to return to his long-abandoned, yet ground-breaking, Ph.D. work and also brought Tesseract into the open source domain. In so doing, Google helped to spur further research interest into efficient and accurate document recognition¹. In the roughly eight years since the project was revived, support has been added for recognition of over fifty languages. Advanced page layout analysis techniques have been implemented in order to detect various types of documents ranging from novels, magazines, newspapers, images, textbooks, sheet music, etc. Language and script detection modules have also been implemented in order to autonomously determine what processing should be carried out for any given world document [6]. If Google's endeavor is successful, then the resulting implications to society will be extraordinary, possibly similar to the impact that Arabic scholars had on Europe when sharing and

¹ The term, recognition, is herein used to describe a machine's extraction of a document's contents. This requires both the document page layout analysis as well as algorithms which subsequently convert the page layout contents into a machine-understandable form. The field of document layout analysis is further discussed in Section 1.2.

1 Introduction

translating ancient Greek literature. If Google is successful in the autonomous digitization and recognition of any printed document regardless of its origin, then it will not be long before information from all of the world's documents become instantly accessible in every language and to everyone around the world. Such a development would certainly speed up the world's already significant progress toward an era of far greater enlightenment and wisdom than has yet been seen.

The autonomous recognition of all printed documents would not only expedite the global advancement of knowledge and wisdom, but would also have tremendous implications toward every individual in society. Such a breakthrough would be especially significant toward the endeavor of Assistive Technology. With many devices being developed and studies being carried out on ways to enhance human computer interaction (HCI) for visually or physically handicapped individuals, digital access to all printed documents could make finding information, not only more convenient, but also possible for many who would not otherwise have access. Global autonomous document recognition could also help open the doors toward breaking down language barriers in information accessibility.

As research and development continues to enhance the accurate translation of discourse between various languages [7], the successful recognition of printed documents could eventually allow them to be machine-translated according to the language preference of a given user. With instant access to all of the world's information, regardless of its language or origin, at one's disposal, collaboration and learning among individuals across the world will be significantly enhanced. All people in the world regardless of their language preference, geographical location, and physical ability will have access to the world's stores of knowledge, and the opportunity to have a profound impact on society through the medium of the World Wide Web. Enhanced document analysis and recognition capabilities will make a significant contribution toward this end. The following section will discuss the background as well as some of the fundamental problems faced in the fields of document analysis and recognition.

1 Introduction

1.2 Introduction to OCR and Document Analysis: A Brief History

From Herbert Shantz's *History of Optical Character Recognition (OCR)* [3], it is clear that the OCR of printed documents has been studied extensively over the last century. In one of the earliest OCR patents [8] (Figure 1), a mechanical apparatus was used to measure the incidence of light reflected back from a printed character when illuminated through a set of character templates. A character detection would occur when the light emitted from the template overlapped the character (assumed to be in dark print) sufficiently to prevent light from being reflected upon the medium. Despite

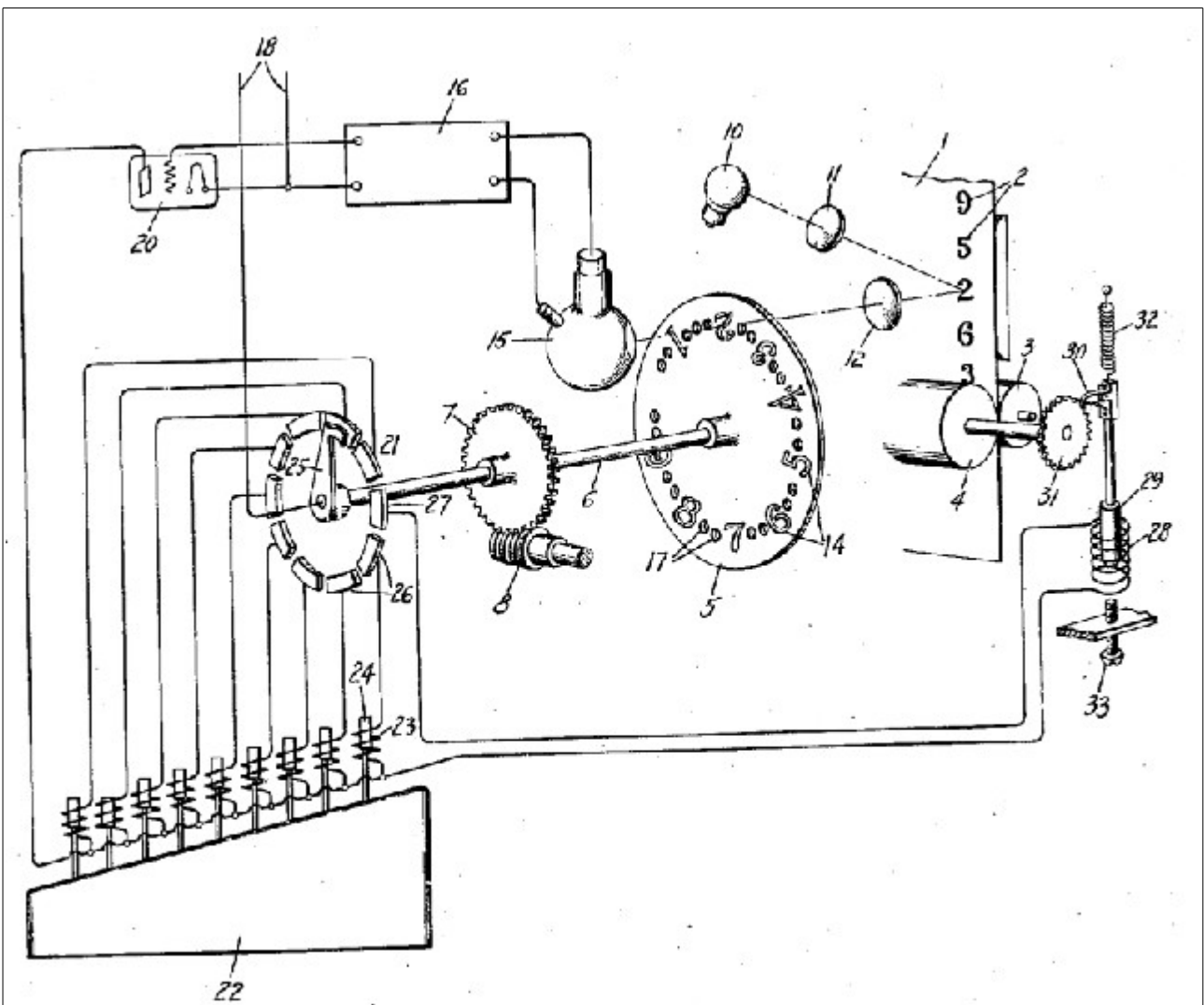


Figure 1: An illustration taken from the 1933 patent from Paul W. Handel, a former employee of General Electric, entitled "Statistical Machine" [8]. This is one of the earliest OCR devices ever invented.

1 Introduction

requiring a significant amount of human intervention to ensure proper alignment and being largely inefficient at best, the fundamental ideas which motivated this early initiative are seen repeatedly throughout the century, and even now, albeit on a much larger scale.

Although some of the first commercial OCR systems were released during the 1950's, their applicability was limited in that, by and large, they were only capable of handling a single font type with very strict rules on character spacing. It was not until the mid-late 1970's, with the invention of both the charge-coupled device (CCD) flatbed scanner and the "Kurzweil Reading Machine" [9] that it became possible for a computer to read a variety of documents with reasonable accuracy. Although the training process for a particular font would take several hours and multi-column page layouts or images had to be specified by the user manually, Kurzweil's software showed significant improvement over the state-of-the-art technologies of the time.

In the 1980's, a company called Calera Recognition Systems [10] introduced an omnifont system that could read pages containing a mixture of fonts while also locating pictures and columns of text without any user intervention or extra training. The progress of the state-of-the-art in document recognition will be discussed from this point in history forward in Chapter 2. More recent commercial OCR systems such as ABBYY FineReader [11], OmniPage Professional [12], and Readiris [13], are all quite accurate, not only in recognizing individual words or characters, but also in understanding and reproducing document layout structure. A magazine or newspaper page may, for instance, contain an intricate heading structure followed by multiple columns of text, pull-out quotes, in-set images, and/or graphs as demonstrated in the historical New York Times article shown in Figure 2 [14].

In order to understand and recognize content of such a document, it is essential to first carry out document layout analysis techniques which will determine how the document is partitioned. The text will be recognized with an understanding of where the columns of text are, which portions of text indicate headings or quotes, and which segments correspond to images, tables, captions, etc. If the text is not partitioned appropriately prior to recognition then the textual output will become unpredictable. With columns, paragraphs, or other structures merged together incorrectly, the text will lose much of its intended meaning and become far less readable to the human eye. For these reasons, sophisticated page layout analysis algorithms are of the utmost importance, not only for document recognition accuracy, but also in ensuring that the generated output is formatted correctly.

1.3 Google's Open Source Initiative

There are various languages, dialects, and page layout formats for which Google's Tesseract software is being developed. Among them are mathematical equations, tables, graphs, and other figures which can be found in any standard science or math text book. While Smith's original work was optimized solely for the recognition of English newspaper formats, Google's continued efforts are aimed at recognizing page formats from a much broader scope [5]. Much of Google's ideas regarding document recognition are essentially in their infancy, and have a long way to go before being fully realized. Although an experimental equation detector has been added to the Tesseract software, its results, although showing significant promise, have been tested to have fairly limited accuracy. A table detector implemented by Google has also been tested on some sample images [16] (Figure 3) to show that, it too, could use significant improvement (Figure 4). Notice that, in the left-most table in Figure 4, the software failed to indicate the years as either belonging to the table or the normal text. They were simply disregarded. Also, the software was unable to determine where exactly the table boundaries are (which should be labeled green). In the right-most table, notice that although a better job was done, while the bottom portion of the text consists of footnotes, it is therein incorrectly labeled as part of the table. Also, the second line of all column labels are not recognized as part of the table when they clearly should be.

The problem of efficiently and accurately detecting equations, tables, graphs, and other figures for the broad spectrum of possible document types is certainly no easy one to solve. Although from a human's perspective, this problem may seem trivial, programming a machine to sum up a document with the same accuracy as the human eye proves to be a daunting task, as will be further discussed in the literature review chapter of this paper. As the inventors of Google continue to work toward their dream of creating an online "Library of Alexandria," there is significant progress to be made before such a large-scale endeavor can be fully realized. The Google Book Search initiative has opened up many avenues for future research in document understanding and recognition, of which, this project is certainly one of the many to come.

1 Introduction

	Total acreage under corn crops.	Total acreage under wheat only.
1870	7,570,379	3,247,873
1875	7,528,543	3,198,547
1880	6,993,699	2,746,733
1885	6,569,105	2,349,305
1890	6,281,494	2,255,894
1895	5,718,997	1,339,806
1898	6,731,463	1,987,385

Almost the entire reduction in the acreage under corn crops, it will be seen, must be due to the reduction of the acreage under wheat, which is a great and conspicuous fact, implying remarkable changes in the economic and political condition of the country. Similarly, there has been an increase of the acreage

CATTLE.

	Millions in or about 1870.	Millions in or about 1898.
United Kingdom	9.2	11.1
France	11.3	13.4
Germany	15.6	18.5 ⁴
Austria	7.4	8.6 ⁵
Hungary	5.3	6.7 ⁶
Italy	3.5 ¹	5.0 ⁵
Belgium	1.2 ²	1.4 ⁶
Holland	1.4	1.6
Denmark	1.2	1.7
Sweden	2.0	2.6
Norway	1.0 ³	1.0 ⁷
Russia in Europe (excluding Poland).	21.4	32.9 ⁷
Total	80.7	104.5

¹ 1875-6. ² 1886. ³ 1895. ⁴ 1897.
⁵ 1890. ⁶ 1895. ⁷ 1900.

Figure 3: The above text includes excerpts from two different pages taken from a scan of Sir Robert Griffin's Stastics textbook (circa 1913) [12]. On the left is a table followed by a paragraph of text, while on the right is a larger table. These images were extracted from a PDF which was made available online by Google.

crops, and the total acreage under wheat in particular, were diminished as follows : ing of a set of influences upon agriculture generally which affects all the old countries of Europe :

	Total acreage under corn crops.	Total acreage under wheat only.
1870	7,570,379	3,247,873
1875	7,528,543	3,198,547
1880	6,993,699	2,746,733
1885	6,569,105	2,349,305
1890	6,281,494	2,255,894
1895	5,718,997	1,339,806
1898	6,731,463	1,987,385

Almost the entire reduction in the acreage under corn crops, it will be seen, must be due to the reduction of the acreage under wheat, which is a great and conspicuous fact, implying remarkable changes in the economic and political condition of the country. Similarly, there has been an increase of the acreage

	Millions in or about 1870.	Millions in or about 1898.
United Kingdom	9.2	11.1
France	11.3	13.4
Germany	15.6	18.5 ⁴
Austria	7.4	8.6 ⁵
Hungary	5.3	6.7 ⁶
Italy	3.5 ¹	5.0 ⁵
Belgium	1.2 ²	1.4 ⁶
Holland	1.4	1.6
Denmark	1.2	1.7
Sweden	2.0	2.6
Norway	1.0 ³	1.0 ⁷
Russia in Europe (excluding Poland).	21.4	32.9 ⁷
Total	80.7	104.5

¹ 1875-6. ² 1886. ³ 1895. ⁴ 1897.
⁵ 1890. ⁶ 1895. ⁷ 1900.

Figure 4: Above is the text from Figure 3 after having been labeled by Tesseract's table detection software. The text within the blue rectangles was identified as belonging to a table while the text within red rectangles was not. The green rectangle should encompass the entire table figure. As can be seen there are both false negatives and false positives.

1.4 Contributions of this Thesis

This thesis introduces a novel approach to detecting mathematical expressions during the document layout analysis stage of OCR. By utilizing and interfacing with the existing data structures and algorithms present within Google's open source OCR engine, Tesseract, much of the more well-studied areas of OCR / document analysis research are surpassed so that a study of the relevant problem of equation detection can be explored in much greater detail than would be possible otherwise. As the Tesseract software, much like commercial state-of-the-art systems, is capable of partitioning a document into columns, paragraphs, headings, etc., the software implemented in this work searches Tesseract's resulting partitions in order to detect regions of interest. Greater document understanding is accomplished through recognition of a variety of relevant features, many of which have yet to have been explored in existing research. Relevant features are subsequently combined with a binary support vector machine (SVM) classifier.

The feature recognition and classification steps in the proposed system are carried out in two separate passes: the first of which detects areas of interest at the individual character level while the second uses results from the first to combine the equation regions into their full partitions. The contributions of this work are summarized in the list below. All of the subjects for which the author could not find an in-depth previous study involving equation detection are marked with an asterisk (*).

- (*) Generation of an extensive ground-truth training set of equation test data, taken from scientific text books and articles. These publications are all in the public domain and thus will be freely available online for future research endeavors.
- Recognition of the following combination of features on equation detection. Although many of these features have indeed been studied to some extent, they have yet to all be used within a single framework as of this date to the author's knowledge.
 - (*) Measurement of the number of horizontally adjacent characters to the right of a given character within some vertical threshold (see chapter x.x).
 - Sub-script/super-script recognition

1 Introduction

- (*) Running a Tesseract classifier trained using Infty Reader's database of mathematical expressions and comparing the result to normal language output to detect math expressions (possibly subsequent recognition?) (see chapter x.x)
- Use of n-grams to locate expressions embedded within text (see chapter x.x)
- Testing whether or not a character's language² classification result falls under a category of potential math symbols such as <, >, _, +, -, /, %, etc.
- Vertical distance to nearest character neighbor above and below (within a horizontal threshold), horizontal distance to nearest neighbor left and right (within some vertical threshold).
- Ratio of horizontal and vertical distance from nearest neighbor to the left and above to nearest neighbor to the right and below respectively.
- Height of characters as compared to average character height within a page
- Detection of Italicized text
- Horizontal bar detection
- Features extracted from PDF if available
- Detection of indentation or centering of text
- Measurement of character vertical distance from a row of text's baseline
- (*) Use of a Support Vector Machine (SVM) classifier for equation detection
- Thorough evaluation of the proposed system's results, includes a comparison with Google's system.

1.5 Organization of Thesis

The work to be discussed in this thesis is aimed toward moving the world a step closer to realizing some of the lofty goals set by Google's engineers and scientists. Chapter 2 presents a review of existing document analysis techniques with extra emphasis on those involving mathematical/scientific documents. Although there are a wide variety of problems which need to be tackled in the area of document recognition, the primary focus is on enhancing equation detection accuracy through the use of feature recognition and a support vector machine (SVM) classifier. The

² Here the language classification result indicates the result of a classifier that was trained for a particular language. Although in the context of this work English is all that is tested, testing of existing techniques in various languages is encouraged for future work.

1 Introduction

remainder of this thesis is organized as follows: Chapter 3 consists of a theory section discussing the image processing and classification techniques employed as well as software optimization techniques utilized by Google's open source OCR engine, Tesseract. Chapter 4, the method section, discusses the ground truth generation procedure, feature recognition algorithms, classification technique, and result evaluation. Chapter 5, the results section, will involve a discussion of all results and their significance. Chapter 6, the conclusion, summarizes important points and discusses recommendations for future work.

2 Literature Review

2.1 Text styles

2.2 Text styles

Heading 4: This is a paragraph heading

Heading 5: This is a sub-paragraph heading

Heading 6: This is a sub-sub-paragraph heading

If you need more headings within the chapters of the document, use Heading 5 or, if necessary, Heading 4 and Heading 5 together.

Paragraph inline headings. An additional possibility is to use paragraph inline headings. Inline headings are emphasized titles at the beginning of a paragraph. It has been tried and tested to provide more clarity and navigability in a lengthy text with these inline headings.

Paragraphs without an inline header are continuations of paragraphs with an inline header. They use a different paragraph style, lacking indentation to support the continuation impression.

Paragraph verbatim inline heading. You would use this verbatim inline heading style for source fragments or verbatim quotes of computer-generated text such as filenames etc..

Emphasis and Deemphasis. It is quite usual to have a style for *emphasizing* text. In this template there is also a style for *deemphasizing* text. You might e.g. adopt the custom to deemphasize the words “et al.” when referring to multiple authors: John Curloe *et al.*, for example.

Mathematical formulas. You might use an
$$E=mc^2$$
 where applicable or a stand-alone formula with its own numbering. Choose one of the dedicated paragraph styles for either right alignment of formulas:

(1)

or for centered alignment of formulas:

(2)

Source code. To place small chunks of source code or verbatim quotes of computer-generated text inline into your text, use this inline style for source and verbatim text. For whole paragraphs of source or computer-generated text, use the dedicated paragraph style for pre-formatted text:

<source code>

2 Literature Review

<source code>
<source code>
<source code>
<source code>

Refer to ► appendix B (p. III) for a description how to place page-long listings of source code into your document together with source highlighting.

Draft mode. There are some text styles for special purposes. For example, while developing a thesis it is handy to mark paragraphs as “in draft quality”. For that purpose, two paragraph styles are provided, one for paragraphs with inline headers and one for those without them:

Draft mode paragraph. This is a paragraph style for draft quality paragraphs with inline heading.

This is a paragraph style for draft quality paragraphs without inline heading.

Todo items. While developing a thesis you will encounter the need to place todo items within your text. They should be marked out to be easily recognized later on. For small inline todo notes and marks use the **inline todo style**, for whole paragraphs use the dedicated paragraph style:

This is the paragraph style for todo items.

As you see, consecutive todo item paragraphs are joined together.

1. It is also possible to mark numbered list items as todo items.
2. It is also possible to mark numbered list items as todo items.

- It is also possible to mark bulleted list items as todo items.
- It is also possible to mark bulleted list items as todo items.

1.1 Special text styles for patterns

Summary. Patterns are a special form of verbalizing content in computer science. The rest of this sub-chapter contains some paragraph styles that you can use to format a pattern. The subpart titles are chosen with reference to the PLML pattern format. The pattern title would appear in the chapter or sub-chapter that is reserved to contain the pattern.

<pattern synopsis>

Problem. Problem description.

2 Literature Review

Context. Context description.

Solution. Solution description.

Inline heading. Another paragraph of the solution description, with inline heading.

And another paragraph of the solution description, without inline heading.

Evidence: Rationale. <rationale description>

Related patterns.

- **Related Pattern.** Related pattern description.

- **Related Pattern.** Related pattern description.

1.1 Lists

Numbered lists. You might use numbered lists together with inline headings:

1. **List item 1 line header.** List item 1 text.
 1. **List item 1.1 line header.** List item 1.1 text.
 2. **List item 1.2 line header.** List item 1.2 text.
2. **List item 2 line header.** List item 2 text.
3. **List item 3 line header.** List item 3 text.

It is however no problem to do without inline headings of course. But remember to right-click the paragraph and choose “restart numbering”.

1. List item 1 text.
2. List item 2 text.
3. List item 3 text.

Bulleted lists. You might use bulleted lists together with inline headings or without them or in mixed form:

- **List item line header.** List item text.

- ☐ List item text.
- ☐ List item text.

2 Literature Review

- List item text.

- **List item line header.** List item text.

Note the fully worked out hierarchy of bullets of this bullet list style:

- List item level 1.

- List item level 2.

- ◆ List item level 3.

- ◇ List item level 4.

- ✦ List item level 5.

- ◇ List item level 6.

- List item level 7.

- List item level 8.

- List item level 9.

- ↪ List item level 10.

Definition lists. The last list style available is the “definition list”. Something similar is known from LaTeX and comes in very handy there:

<definition list term 1>

<definition list description 1>

<definition list term 2>

<definition list description 2>

1.1 Linking and referencing

URLs. You might use footnotes to mention URLs directly and not via bibliographic references, e.g. to mention the `example.org` site³. This does not clutter the text with URLs but is better than linking without mentioning the URL, as it preserves full functionality for printed versions.

³<http://www.example.org>

2 Literature Review

Footnotes. And you might use footnotes for additional annotations⁴ that have no place in the flow of thoughts. As you see, we place a special character in front of every footnote to mark out hyperlinks in PDF versions better.

Internal references. All hyperlinks (including document-internal references) are prepended with a link flag in this template. Link flags help detecting active elements in PDF documents but can become ugly if there are too much. The following basic styles are proven:

- (see ►p. 17)
- see ►object 1 (p. 21)
 - see ►chapter Error: Reference source not found (p. Error: Reference source not found)
 - see ►appendix B (p. III)
- (see ►object 1, p. 21)
 - (see ►chapter Error: Reference source not found, p. Error: Reference source not found)
 - (see ►appendix B, p. III)
- do not prepend glossary entries with a link flag

1.1 Citing and bibliography

Citing. Block quotes have their dedicated paragraph style and might span one or multiple paragraphs:

This is a paragraph where some other work is cited. Which means that this very text that you read is the tet of the citation, drawn from this very other work. For convenience, an unknown dummy work is cited.
Another paragraph of the citation is appended using a forced linebreak, not by starting a real new paragraph. ►[p. 1-198]

Bibliographic references Here are examples of bibliographic references of every type used in this template. See the bibliography and meta pages on how these bibliographic types differ. Note that these bibliographic references are hyperlinked in the PDF output though this is not natively supported by OpenOffice.org yet. The idea is to mark bibliography items as headings (menu “Extras :: Chapter numbering ...”), then

⁴Such as this annotation.

2 Literature Review

insert hyperlinks to headings with the bibliographic reference as link text. Do this just before finishing your document or, even better, implement it in OOO or bibus.

- ARTICLE: ►[, pp. 199-201₋₁₉₈]
- BOOK: ►[, pp. 1.3]
- INBOOK: ►[, p. 543₊₁₂₀]
- INCOLLECTION: ►[]
- INPROCEEDINGS: ►[]
- MASTERTHESIS: ►[]
- MISC: ►[]
- PHDTHESIS: ►]
- WWW: ►, ► chp. 4.1] (a bibliographic reference including a hyperlinked subpart marker)

1 Object demonstration

Summary. In this template, all framed content is referred to as “objects” regardless of the actual content (images, tables, diagrams etc.). So only one index of objects is necessary, which is far more clear than one index for each type of frames. Note that all frames are anchored to the paragraph whose text starts *below* the frame.

Table object. Here is a demonstration of a table within a frame. Note the additional OOo Draw elements placed over the table and anchored to the frame. For graphical tables such as this better use hard formatting than paragraph styles, to not clutter your style namespace.

(01) text	(08) text	(15) text	(22) text	title 1
(02) text	(09) text	(16) text	(23) text	title 2
(03) text	(10) text	(17) text	(24) text	title 3
(04) text	(11) text	(18) text	(25) text	title 4
(05) text	(12) text	(19) text	(26) text	title 5
(06) text	(13) text	(20) text	(27) text	title 6
(07) text	(14) text	(21) text	(28) text	title 7
title 8	title 9	title 10	title 11	
title 9/10				

no focus partial focus focus

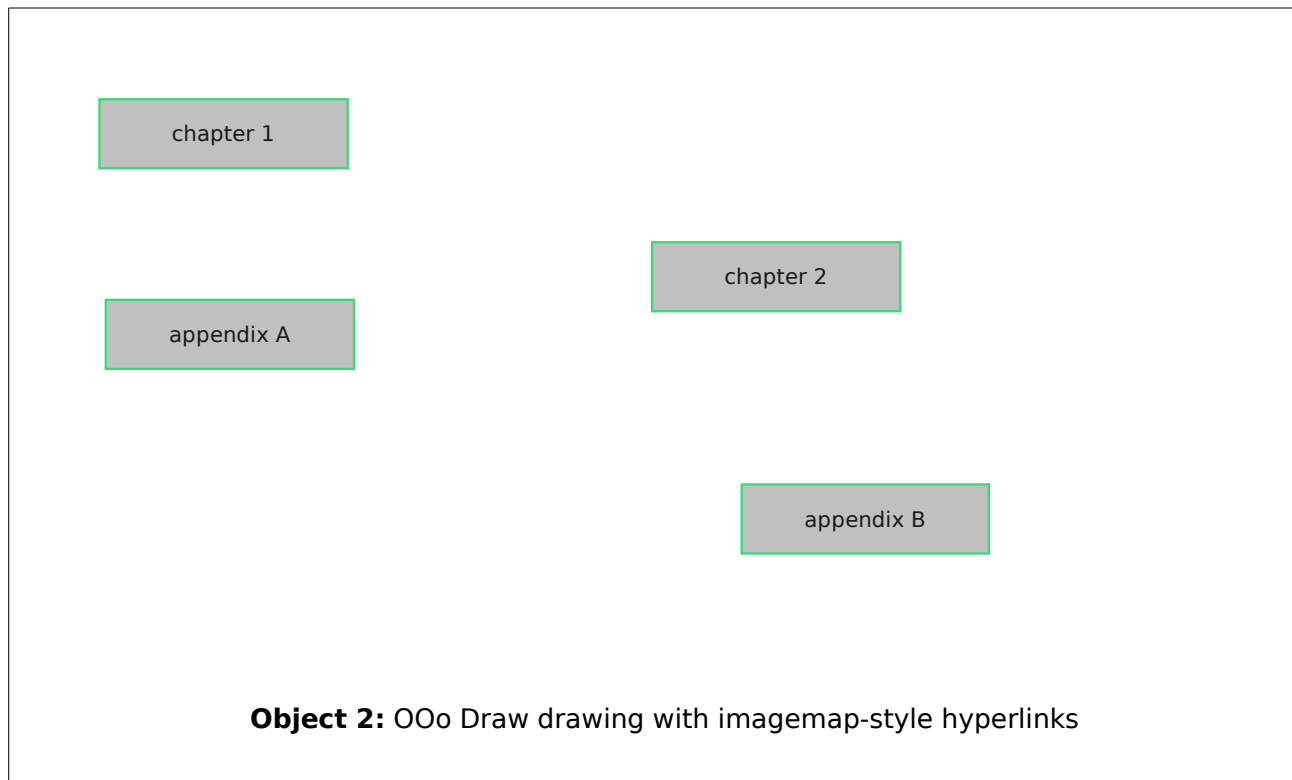
Object 1: table with OOo Draw elements

OOo Draw drawing with imagemap-style hyperlinks as object. The only way to include vector-oriented graphics in your document is to include OOo Draw objects. Now using OLE-objects for that purpose imposes cumbersome editing, placing and adjusting. Using the OOo Draw-like functionality of writer lacks Draw styles. The solution is to draw with styles in OOo Draw, group the whole drawing and then paste it into a frame here in OOo writer. This was done in the following example.

1 Object demonstration

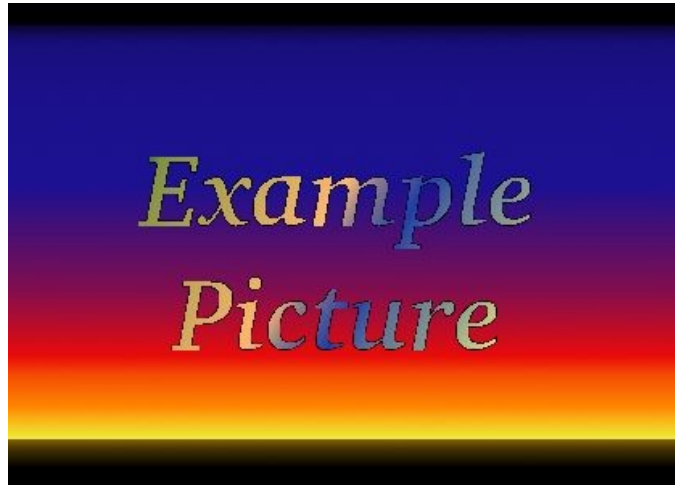
Another goodie of this example is that it shows a possibility to create imagemap-style hyperlinks within OOO Draw diagrams that are *usable* in the exported PDF documents. The green boxes are transparent non-printing copies of one PNG image which have been placed over the inserted OOO Draw diagram and are anchored to the frame. They can be hyperlinked to outline elements of this document or to other targets.

Note further that it is a good idea to place a white image into the background of the frame, which is aligned to the frame. This is to span the desired size of the frame, so lets you work around frame sizing problems and problems with the placement of the frame's title.



Picture object. This one is easy: a picture within a frame.

1 Object demonstration



Object 3: framed picture demonstration

A Glossary of terms and abbreviations

Summary. There have been special paragraph styles defined for the glossary. Do not use the styles for definition lists because the glossary styles have been adjusted to appear as PDF bookmarks in PDF documents exported from OpenOffice.org.

glossary term 1

This is the definition and explanation of glossary term 1.

glossary term 2

This is the definition and explanation of glossary term 1. The indentation of citations has been adjusted so that you can reasonable use it within the glossary, too:

Citation text. Citation text. Citation text. Citation text. Citation text.
Citation text. Citation text. Citation text. Citation text. Citation text.
Citation text. Citation text. Citation text. ►[, p.]

B Source listings

Summary. This appendix chapter will contain source codes developed during the diploma thesis, ordered by program modules. If you want to include long source listings here it might be a good idea to use source highlighting. The best way seems to use an editor which can export highlighted source to HTML (such as KDE's kate), to open the HTML document with OOo writer and then to copy it into your thesis. This does result in hard formatting (not style-based) but this does not hurt here. A short example done with this method is included here.

B.1 http_post()

<?php

```
// adapt these constants and variables to configure the script
// include the path of commands if they reside outside of PHP's PATH
define('LOGFILE_NAME', __FILE__ . '.log.txt');

/** perform a HTTP POST request using the cURL PHP extension
 * @param $server where to POST to, e.g. http://www.example.org
 * @param $path URL part after server name, e.g. '/foo/bar.php'
 * @param $vars array of key/value pairs, maybe nested; or an object
 * @return the content returned by the server, without headers
 */
function http_post($server, $path, $vars) {
    $ch = curl_init();
    curl_setopt($ch, CURLOPT_URL, $server.$path);
    curl_setopt($ch, CURLOPT_POST, 1);
    curl_setopt($ch, CURLOPT_POSTFIELDS, http_build_query($vars));
    curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);
    $reply = curl_exec($ch);
    if (curl_errno($ch))
        error_log(
            "ERROR: curl_exec() error no ". curl_errno($ch) . " : " .
            curl_error($ch)."\n",
            3, LOGFILE_NAME
        );
    curl_close($ch);
    return($reply);
}

?>
```

Index of glossary items

► glossary term 1 I | ► glossary term 2 I

Index of objects

- ▶ Object 1: table with OOo Draw elements **7**
- ▶ Object 2: OOo Draw drawing with imagemap-style
hyperlinks **8**
- ▶ Object 3: framed picture demonstration **9**

Bibliography

Bibliography

[1] K. Wilcox and A. Stephen, "*Are Close Friends the Enemy? Online Social Networks, Self-Esteem, and Self-Control*," Journal of Consumer Research, Forthcoming Columbia Business School Research Paper No. 12-57, Date posted: October 3, 2012.

[2] D. A. Vise and M. Malseed, *The Google Story*. New York City: Dell Publishing, 2005.

[3] H. F. Shantz, *The History of OCR, Optical Character Recognition*. Manchester Center: Recognition Technologies Users Association, 1982.

[4] S. V. Rice, F. R. Jenkins and T. A. Nartker, "*The Fourth Annual Test of OCR Accuracy*," Technical Report 95-03, Information Science Research Institute, University of Nevada, Las Vegas, July 1995.

[5] L. Vincent, "Google Book Search: Document Understanding on a Massive Scale," *International Conference on Document Analysis and Recognition (ICDAR)*, 2007, pp. 819 - 823.

[6] R. Unnikrishnan and R. Smith, "*Combined Script and Page Orientation Estimation Using the Tesseract OCR Engine*," Submitted to International Workshop of Multilingual OCR, 25th July 2009, Barcelona, Spain.

[7] Z. Huang, M. Cmejrek and B. Zhou, "Soft Syntactic Constraints for Hierarchical Phrase-based Translation Using Latent Syntactic Distributions," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, p. 138-147.

[8] P. W. Handel, "*Statistical Machine*," United States Patent Office. 1,915,993, Jun. 27, 1933.

[9] A. Kleiner and R. Kurzweil, *A Description of the Kurzweil Reading Machine and a Status Report on its Testing and Dissemination*, Bulletin of Prosthetics Research, vol. 27, no. 10, pp. 72-81, Spring. 1977.

Bibliography

[10] M. Bokser, *Omnidocument Technologies*, Proceedings of the IEEE, vol. 80, no. 7, pp. 1066-1078, Jul. 1992.

[11] ABBY FineReader. "ABBY FineReader for Personal Use." Internet: <http://finereader.abby.com/>, Date Accessed: April 3, 2013.

[12] Nuance Inc. "OmniPage Professional." Internet: <http://www.nuance.com/for-business/by-product/omnipage/professional/index.htm>, Date Accessed: April 3 2013.

[13] Iris Products and Technologies. "Introducing the new Readiris 14." Internet: <http://www.irislink.com/c2-2115-189/Readiris-14--OCR-Software--Scan--Convert---Manage-your-Documents-.aspx>, Date Accessed: April 3, 2013.

[14] Contributor: Bob Stein (uploader to <http://archive.org>). "New York Times August September 1901 Collection." Internet: http://archive.org/download/NewYorkTimesAugSept1901Collection/New_York_Times_August_September_1901_Part_7_text.pdf, Date Accessed: March 14, 2013.

[15] T. M. Breuel and U. Kaiserslautern, "The hOCR Microformat for OCR Workflow and Results," *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, 2007, pp. 1063 - 1067.











[16] R. Griffin, *Statistics*. London: Macmillon and Co., 1913, pp. 121-122.

Colophon

<A colophon (literally, end stroke) is an inscription at the end of a written work, containing facts about its production. It may name artists, printers etc. and discuss typographic and technical details such as typefaces and papers.>

Attached electronic data

Description of attached files and folders⁵

| | | |
|---|---------------|---------------|
|  | <file name> | <description> |
|  | <folder name> | <description> |
|  | <file name> | <description> |
|  | <folder name> | <description> |
|  | <folder name> | <description> |
|  | <file name> | <description> |
|  | <folder name> | <description> |
|  | <file name> | <description> |
|  | <file name> | <description> |
|  | <file name> | <description> |
|  | <folder name> | <description> |

⁵You might find additional auxiliary files for download at this thesis' web location.

Accessing the attached electronic data

The paper version of this thesis should contain an envelope with an optical data medium (CD or DVD) here.

The digital version of this thesis contains the content of this medium as attachments to the PDF file. If your PDF viewer cannot handle attachments you may access these files at this thesis' web location.

