UNIVERSITAT POLITÈCNICA DE VALÈNCIA
INSTITUTO TECNOLÓGICO DE INFORMÁTICA

PATTERN RECOGNITION AND
HUMAN LANGUAGE TECHNOLOGY GROUP

IARFID Master Thesis

Interactive Layout Analysis

Author: Lorenzo Quirós Díaz

Advisor: Carlos D. Martínez Hinarejos
Co-advisor: Alejandro Héctor Toselli
Co-advisor: Enrique Vidal Ruiz

September 5, 2016

Hi, nothing there yet :) ...

Acknowledgements

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Chapter Outline

1.1 Introduction

- Whats HTR and Layout Analysis
- Whats Interactive Pattern Recognition
- What are we going to do here :)
- Work structure (ie paper estructure)

Several manuscripts have being digitized in order to preserve the valued information from the day-by-day deterioration of the physical document. The amount and importance of the information contained on those manuscripts motivates the development of several techniques and tools to explore, analyze and read them in a more comprehensive manner . This is from image quality enhancement citesome Examples here to automatic Handwritten Text Recognition (HTR) citesome Examples here

1.2 Motivation

- Why to focus on Layout Analysis and Interactive Pattern Recognition
-

1.3 Related Work

- Image Segmentation
- Heuristics
- HMMs, NN, grammars ....

1.4 Overview of the Proposal Approach

1.5 Context of Applications and Assumptions

1.6 Expected Outcomes/Results

Bibliography

[1] Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

CHAPTER 2

FUNDAMENTS

Chapter Outline

2.1 Fundaments

2.2 Image Segmentation

2.3 Conditional Random Fields

2.3.1 Definition

2.3.2 CRFsuit Toolkit

This is the reference [1]

2.4 Interactive Pattern Recognition

2.5 Gradient Descent

Bibliography

[1] Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

CHAPTER 3

INTERACTIVE LAYOUT ANALYSIS

Chapter Outline

3.1 Interactive Layout Analysis

3.2 System Architecture

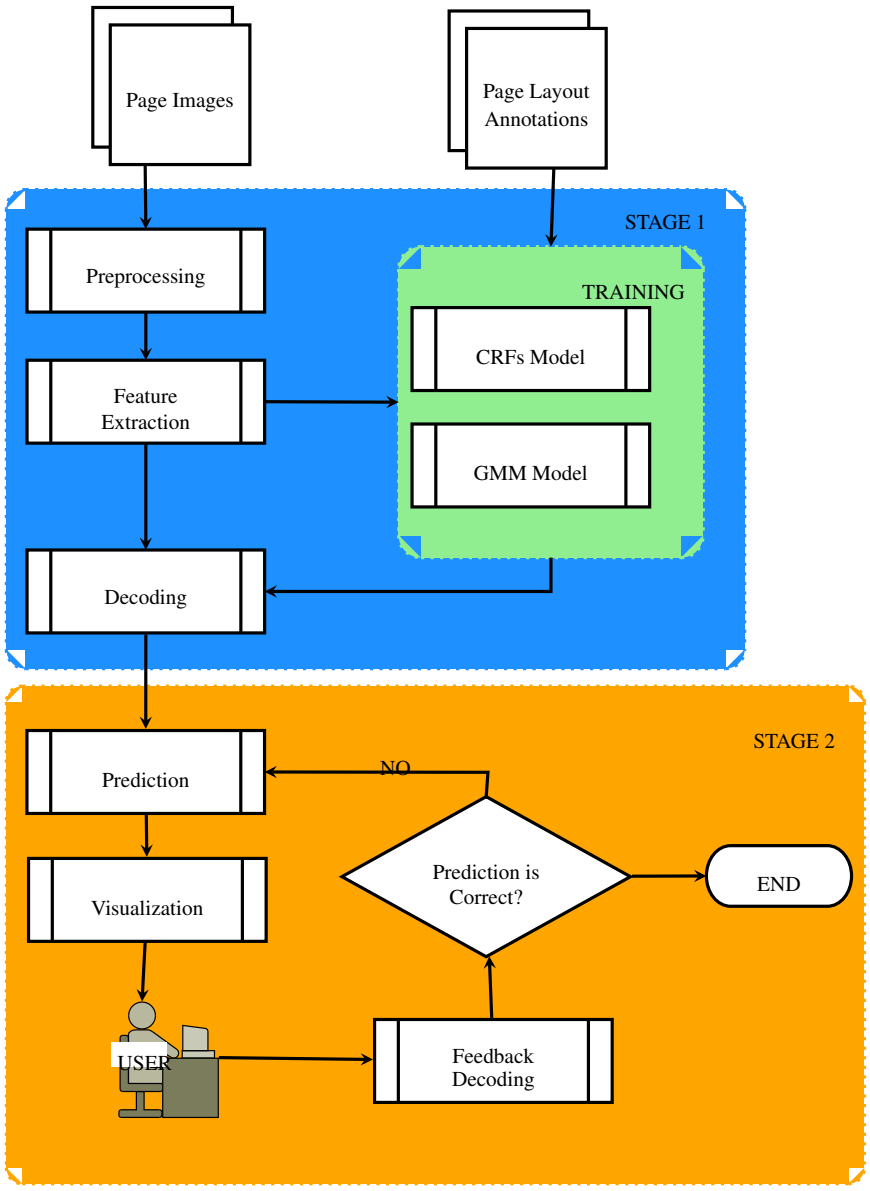


Figure 3.1: System Architecture Diagram

3.3 Preprocessing

3.4 Feature Extraction

3.5 CRF's Learning

3.6 GMM Learning

3.7 Decoding

3.8 User Interaction

3.9 Evaluation Measures

Bibliography

CHAPTER 4

EXPERIMENTS AND RESULTS

Chapter Outline

4.1 Overview

4.2 Corpus Description

To develop and test the new method proposed in this document a manuscript under following requirements is preferred to be selected:

- Reasonably good state of preservation.
- Available in some digital format.
- Reasonably large.
- Well-defined layout zones.
- Layout should be complex enough to exemplify method strengths and weaknesses.
- Ground trough available or easy to build.

Under those constrains, the manuscript chosen for the present work is the first tome of a seven volume manuscript entitled "Historia de las Plantas", PLANTAS for short, a XVII century handwritten botanical specimen book compiled by Bernardo Cienfuegos, one of the most outstanding Spanish botanists in the XVII century. The first volume has 49 pages at the beginning comprising indices, reference tables, a botanical glossary in different languages, and a 36-page preface written by Cienfuegos. This is followed by 887 numbered pages that contain 152 chapters about cereals and related plants, including 126 botanical illustrations. All in all, the first volume has 1 035 pages, containing about 20000 handwritten text lines. This corpus is already digitized at 300ppi in 24 bit RGB color, available as JPG images along with their respective ground trough layout in PAGE XML format [5] compiled by PRHLT group [1] using seven categories, namely: catch-word, heading, marginalia, page-number, paragraph, signature-mark, float (illustrations); see Figure 4.1 for reference.

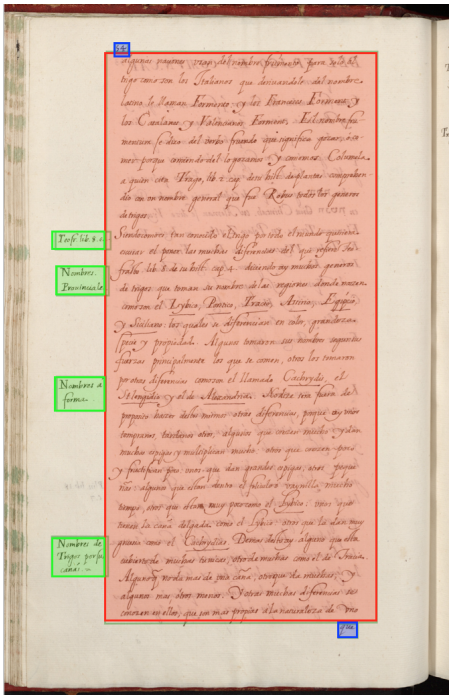


Figure 4.1: PLANTAS Layout Zones; green for marginalia, red for paragraph and blue for catch-word

In this stage of the presented work only a sub-set of 39 pages was considered, in order to complain with stage and time restrictions. Pages that contains indice, reference tables and illustrations were excluded since goals of this stage are restricted to the main paragraph. 22 of those pages were selected for training the model, and the remaining 17 for test (see 5.1 for reference).

4.3 Implementation Notes

System was implemented mainly in Python 2.7 due the facilities provided by the language (N-dimensional array structures, image processing tools, plotting tools, etc) and how quickly is possible to develop a new peace of software. Besides, CRFSuite is used to handle CRF training and tagging steps. On this chapter some implementation details will be explained for clarification, most of them are related to non-direct implementation of the theory explained on Chapters 2 and 3; and implemented towards code optimization.

4.3.1 Integral Image

Integral Image (or Summed Area Table) is a data structure and algorithm first used by Crow [2] in Computer Graphics and introduced by Viola and Jones [6] to Computer Vision is widely used for quickly and efficiently generating the sum of values in a rectangular subset of a grid. The Integral Image I of an $M \times N$ input image A is defined as a 2D cumulative sum of A:

$$I(x, y) = \sum_{x' \leq x, y' \leq y} A(x', y') \quad x \leq M, y \leq N \tag{4.1}$$

Then, using the integral image any rectangular sum can be computed in four array references [6] (see 4.2), this is four access to the I matrix instead $(x * y)$ access to A using direct approach.

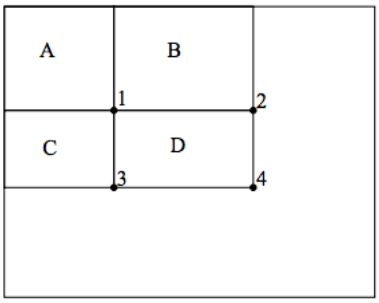


Figure 4.2: The sum of the pixels within rectangle D can be computed with four array references. The value of the Integral Image at location 1 is the sum of the pixels in rectangle A. The value at location 2 is A + B, at location 3 is A + C, and at location 4 is A + B + C + D. The sum within D can be computed as $4 + 1 - (2 + 3)$ [6]

Is correct to copy this image????

Since CRF model provides the probability of each site in the site set S, we can see it as a matrix and use Integral Image for all related calculations. For example on Eq. REF to min function we need to compute the sum of the probability of each site inside each k layout-zone, this can be computed as:

$$\sum_{d=1}^{|S_k|} \log P(s_d|(\boldsymbol{u}_k, \boldsymbol{b}_k)) = I_k(\boldsymbol{b}_k) + I_k(\boldsymbol{u}_k) - I_k(\boldsymbol{u}_{kr}, \boldsymbol{b}_{kc}) - I_k(\boldsymbol{b}_{kr}, \boldsymbol{u}_{kc}) \tag{4.2}$$

4.3.2 Element-wise to matrix evaluation

User interaction requires a system fast enough to keep real time feeling to the user, this is lees than 1.0 seconds [4]. On this system user interacts as explained on 3.8, but, direct implementation of Eq. REF to min function using `for`

loops is just not fast enough to keep real time feeling ($\approx$ 25 seconds per click). In order to reduce time consumed by the CPU to compute Eq. REF to min function a matrix-like version of the equation is implemented, this allows the system to use `numpy` multidimensional and bradcasting features to reduce delay.
Then Eq. REF to min function is transformed to Eq. 4.3, where sums have been expanded to specific the specific case of this work (only main paragraph), I_k meas for the Integral Image of the probabilities of the layout zone k, $P_{\boldsymbol{u}_1}$ and $P_{\boldsymbol{b}_1}$ are the probability matrix from $\boldsymbol{u}_1$ and $\boldsymbol{b}_1$ GMM models respectively, f_r and f_c are the decoded feedback from the user as the row and column selected respectively.

$$
\begin{aligned}
h^*_{\boldsymbol{b}_1} &= \arg\min I_0(-1,-1) \\
&\quad - (I_0(f_r:,f_c:) + I_0(f_r-1,f_c-1) - I_0(f_r-1,f_c:) - I_0(f_r:,f_c-1)) \\
&\quad + (I_1(f_r:,f_c:) + I_1(f_r-1,f_c-1) - I_1(f_r-1,f_c:) - I_1(f_r:,f_c-1)) \\
&\quad + P_{\boldsymbol{u}_1}(f_r,f_c) + P_{\boldsymbol{b}_1}(f_r:,f_c:) \\
h^*_{\boldsymbol{u}_1} &= \arg\min I_0(-1,-1) \\
&\quad - (I_0(f_r,f_c) + I_0(:f_r-1,:f_c-1) - I_0(:f_r-1,f_c) - I_0(f_r,:f_c-1)) \\
&\quad + (I_1(f_r,f_c) + I_1(:f_r-1,:f_c-1) - I_1(:f_r-1,f_c) - I_1(f_r,:f_c-1)) \\
&\quad + P_{\boldsymbol{u}_1}(1:f_r,1:f_c) + P_{\boldsymbol{b}_1}(f_r,f_c)
\end{aligned} \tag{4.3}
$$

Which is the correct notation for this kind of equation, where most of the terms are matrix

Under this approach, time to compute min value is reduced from $\approx 25 seconds$ to $\approx 0.06 seconds$, which is enough for current application.

4.4 Experiments

Experiments have been conducted over selected corpus to obtain model parameters such as features to train CRF models, window size of those features and training algorithm, between others. Parameters search performed is not exhaustive since main goal of this stage is not to get the best model, but to interactive algorithm features.

features?? is this the correct word here???

4.4.1 Conditional Random Fields

CRF performance is highly dependent of features selection, due that, some values of image zoom (Z), window size (W), grid size (G) and the training algorithm (A) have been used to train the CRF model and the best one is selected, results are presented on Table 4.1:

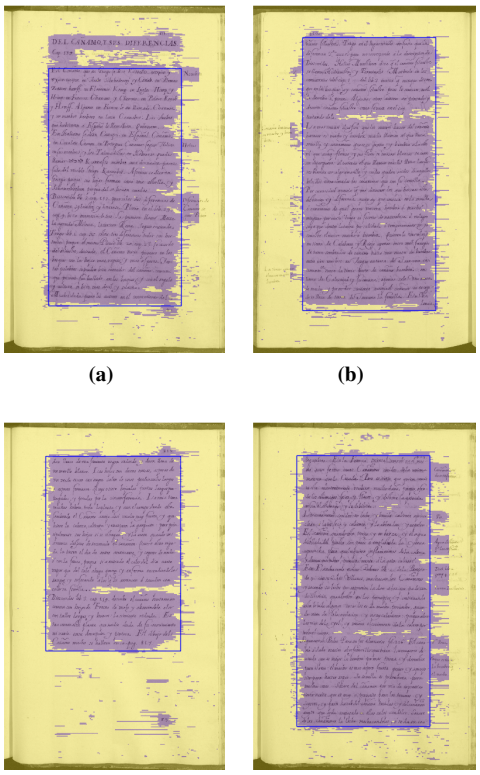
Table 4.1: CRF's parameters search results

	Z	W	G	A	FT[s]	TrT[s]	TeT[s]	P	R	F1
1	0.4	33	12	arow	28.073	19.154	0.998	0.883	0.882	0.879
2	0.4	33	12	lbfgs	28.073	92.451	1.134	0.889	0.884	0.880

Table header explanation: where put it?

Update table using data from param_search, I'm re-computing all the data since I delete the file by mistake :(

Parameters on row XXX of Table 4.1 are selected to train the final model, quantitative results are showed on Table 4.2 and some examples of qualitative results on Figure 4.3



(a) **(b)** **(c)** **(d)**

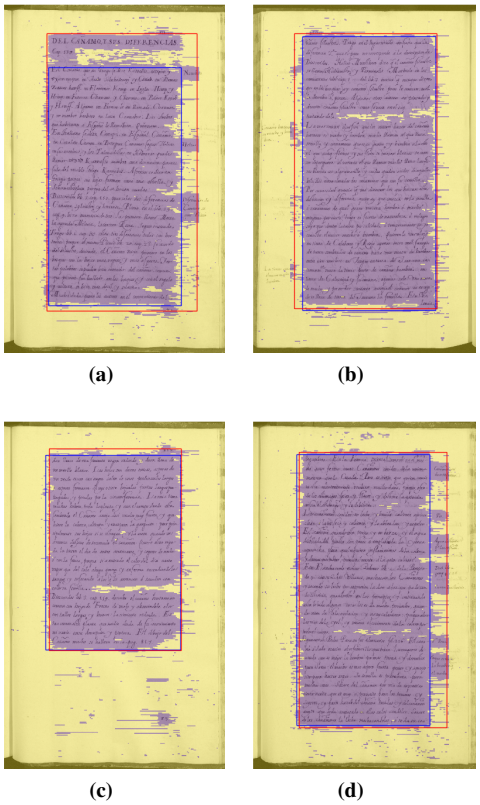
Figure 4.3: CRF's qualitative results, pixels classified as background in yellow, pixels classified as paragraph in purple. Ground through rectangle in blue

Table 4.2: CRF's site level quantitative results

Page	Precision	Recall	F1
0944	0.908	0.914	0.907
0945	0.953	0.953	0.953
0946	0.958	0.956	0.956
0947	0.951	0.952	0.951
0948	0.949	0.951	0.950
0956	0.894	0.896	0.891
0957	0.946	0.945	0.945
0958	0.931	0.927	0.927
0959	0.952	0.952	0.952
0960	0.955	0.954	0.954
0961	0.950	0.950	0.950
0962	0.909	0.921	0.910
0963	0.919	0.914	0.916
0964	0.909	0.910	0.908
0965	0.942	0.942	0.942
0966	0.947	0.947	0.947
0967	0.945	0.946	0.945
Average	0.936	0.936	0.936

4.4.2 Connected Components Labeling (CCL) Approach

Connected components algorithms based on morphological operations are a simple method to detect connected objects or regions in binary images. A simple version of this kind of algorithms [3] is implemented as a point of comparison for proposed method. This is, all adjacent sites classified as "paragraph" by the CRF model are grouped, then we search for the minimum rectangle where all sites of the same group fits and finally, based on user experience, only the biggest rectangle is selected, see Figure 4.4 for some qualitative examples and Table 4.3 for quantitative results.



(a) **(b)** **(c)** **(d)**

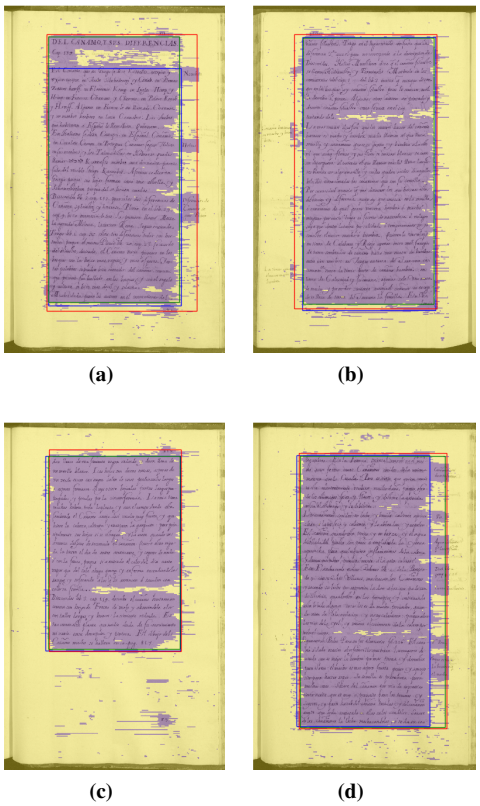
Figure 4.4: CCL results examples, red line. Ground through added for reference, blue line.

Table 4.3: CCL quantitative results

Page	Precision	Goal-Oriented
0944	0.768	0.768
0945	0.926	0.926
0946	0.877	0.877
0947	0.944	0.944
0948	0.938	0.937
0956	0.782	0.782
0957	0.921	0.921
0958	0.864	0.863
0959	0.971	0.971
0960	0.879	0.879
0961	0.942	0.942
0962	0.765	0.765
0963	0.945	0.944
0964	0.790	0.787
0965	0.908	0.908
0966	0.874	0.873
0967	0.912	0.912
Average	0.883	0.882

4.4.3 Prior-Probability Approach

Prior-Probability is used to estimate the best "paragraph" coordinates, this is we maximize Eq. (Add reference!!!) over all $(\boldsymbol{u}_k, \boldsymbol{b}_k)$ in the image range using brute force approach and methods explained in section 4.3. See Figure 4.5 for some qualitative results examples, and Table 4.4 for quantitative results.



(a) **(b)** **(c)** **(d)**

Figure 4.5: Proposed method results example, blue line for ground trough, red line for connected components labeling approach, green line for proposed approach.

Table 4.4: Proposed method quantitative results

Page	Precision	Goal-Oriented
0944	0.865	0.862
0945	0.956	0.956
0946	0.977	0.977
0947	0.943	0.943
0948	0.969	0.969
0956	0.796	0.791
0957	0.960	0.960
0958	0.877	0.875
0959	0.945	0.945
0960	0.978	0.978
0961	0.952	0.952
0962	0.870	0.869
0963	0.888	0.888
0964	0.881	0.878
0965	0.944	0.944
0966	0.952	0.951
0967	0.940	0.940
Average	0.923	0.922

4.4.4 Interactive Approach

User is allowed to change system hypothesis by a simple click over the image, this feedback is decoded an new hypothesis is presented to the user. Number of clicks needed by the user to define the main paragraph and new hypothesis are recorded. Quantitative results are presented on Table 4.5 along with number of clicks performed by an user.

Table 4.5: Post-user-feedback quantitative results

Page	Precision	Goal-Oriented	# Clicks
0944	0.960	0.960	1
0945	0.956	0.956	0
0946	0.977	0.977	0
0947	0.943	0.943	0
0948	0.969	0.969	0
0956	0.936	0.936	2
0957	0.960	0.960	0
0958	0.960	0.960	1
0959	0.945	0.945	0
0960	0.978	0.978	0
0961	0.952	0.952	0
0962	0.961	0.961	1
0963	0.924	0.924	1
0964	0.946	0.946	1
0965	0.944	0.944	0
0966	0.952	0.951	0
0967	0.940	0.940	0
Average	0.953	0.953	—

4.5 Results Discussion

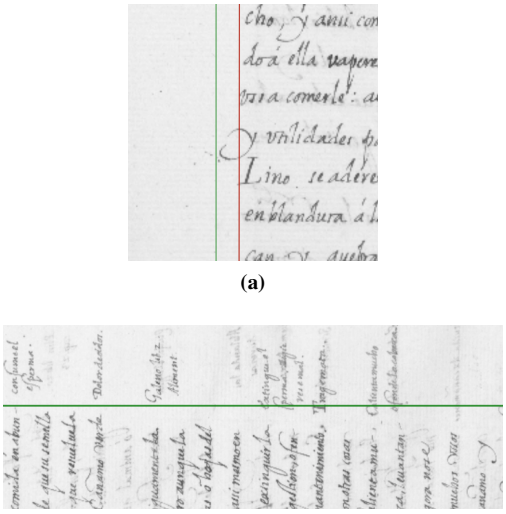
CRF model performed an average of 93.6% F1-score. Despite, the set of features selected were very elemental (only site color intensity and position) and a non-exhaustive parameter search, results are very promising for further stages. Miss-classification of marginalia and title zones could be because color intensity is more a feature to identify text than to identify the different layout zones, that means layout zone classification lays mostly on position feature, also only two classes have been used for training CRF model. Although this model provides posterior probability needed for interactive model, feature selection needs to be improved for further stages of this project.

CCL is highly dependent of geometric distribution of posterior probability from CRF model and, after segmentation, there is no easy and general way to improve results. For this reason this approach is used only as a point of comparison for proposed method. Performance computed by Goal-Oriented and MatchScore is very similar 88.2% and 88.3% respectively (average), which is a good start point, but almost all marginalia and title zones have been classified as paragraph, this is due the high dependence of geometric distribution.

Prior-probability plays a main role in next approach, most of the marginalia zones are not longer classified as paragraph and rectangle boundary is stretched to text boundary which relies on a average 4% improvement over CCL approach, this is up to 13% improvement on complex cases (see Table 4.4).

Finally methods studied on bibliography and the method proposed in this work still having errors that must be fixed manually by some human. Interactive approach based on proposed method provides the framework to help user to fix those errors, under that premise not only the system performance must be computed, but the user effort as well. On this seventeen pages corpus 65% of the times user made no changes to hypothesis provided, thus, hypothesis is

good enough to identify the paragraph. On the other hand 30% of the times user performs only one click in order to fix errors on the hypothesis, finally only in one case two clicks were needed. Performance average is 95.3% on both methods; notice that 100% was not reached in any case because users have a different concept of how much the border of the zone needs to be to the border of the text or cases where two o more blocks cannot be divided by a single line, see examples on Figure 4.6.



(a)

(b)

Figure 4.6: Examples of different zone boundaries provided by different users. a) small section of a character is out due ajust to text border, b) single horizontal line cannot divide upper and bottom zones.

Bibliography

[1] Bosch, V., Bordes-Cabrera, I., Muñoz, P. C., Hernández-Tornero, C., Leiva, L. A., Pastor, M., Romero, V., Toselli, A. H., and Vidal, E. (2014). Computer-assisted Transcription of a Historical Botanical Specimen Book : Organization and Process Overview Categories and Subject Descriptors. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 125–130, Madrid, Spain.

[2] Crow, F. C. (1984). Summed-area tables for texture mapping. *ACM SIGGRAPH Computer Graphics*, 18(3):207–212.

[3] Gonzalez, R. C. and Woods, R. E. (2008). *Digital Image Processing*. Prentice Hall, 3 edition.

[4] Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann Publishers Inc.

[5] Pletschacher, S. and Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-truth Elements) format framework. *Proceedings - International Conference on Pattern Recognition*, pages 257–260.

[6] Viola, P. and Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154.

CHAPTER 5

APPENDIX

Chapter Outline

5.1 Corpus Distribution Table

Table 5.1: Corpus Pages Distribution

Name	ID	used to
Mss_003357_0944_pag-811[843]	0944	test
Mss_003357_0945_pag-812[844]	0945	test
Mss_003357_0946_pag-813[845]	0946	test
Mss_003357_0947_pag-814[846]	0947	test
Mss_003357_0948_pag-815[847]	0948	test
Mss_003357_0956_pag-823[855]	0956	test
Mss_003357_0957_pag-824[856]	0957	test
Mss_003357_0958_pag-825[857]	0958	test
Mss_003357_0959_pag-826[858]	0959	test
Mss_003357_0960_pag-827[859]	0960	test
Mss_003357_0961_pag-828[860]	0961	test
Mss_003357_0962_pag-829[861]	0962	test
Mss_003357_0963_pag-830[862]	0963	test
Mss_003357_0964_pag-831[863]	0964	test
Mss_003357_0965_pag-832[864]	0965	test
Mss_003357_0966_pag-833[865]	0966	test
Mss_003357_0967_pag-834[866]	0967	test
Mss_003357_0158_pag-053[057]	0158	train
Mss_003357_0159_pag-054[058]	0159	train
Mss_003357_0161_pag-056[060]	0161	train
Mss_003357_0162_pag-057[061]	0162	train
Mss_003357_0163_pag-058[062]	0163	train
Mss_003357_0176_pag-071[075]	0176	train
Mss_003357_0177_pag-072[076]	0177	train
Mss_003357_0178_pag-073[077]	0178	train
Mss_003357_0179_pag-074[078]	0179	train
Mss_003357_0180_pag-075[079]	0180	train
Mss_003357_0181_pag-076[080]	0181	train
Mss_003357_0182_pag-077[081]	0182	train
Mss_003357_0183_pag-078[082]	0183	train
Mss_003357_0184_pag-079[083]	0184	train
Mss_003357_0185_pag-080[084]	0185	train
Mss_003357_0186_pag-081[085]	0186	train
Mss_003357_0187_pag-082[086]	0187	train
Mss_003357_0188_pag-083[087]	0188	train
Mss_003357_0189_pag-084[088]	0189	train
Mss_003357_0190_pag-085[089]	0190	train
Mss_003357_0191_pag-086[090]	0191	train
Mss_003357_0192_pag-087[091]	0192	train

NOMENCLATURE

CCL	Connected Components Labeling
CRF	Conditional Random Field
GMM	Gaussian Mixture Model