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

August 10, 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)

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

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

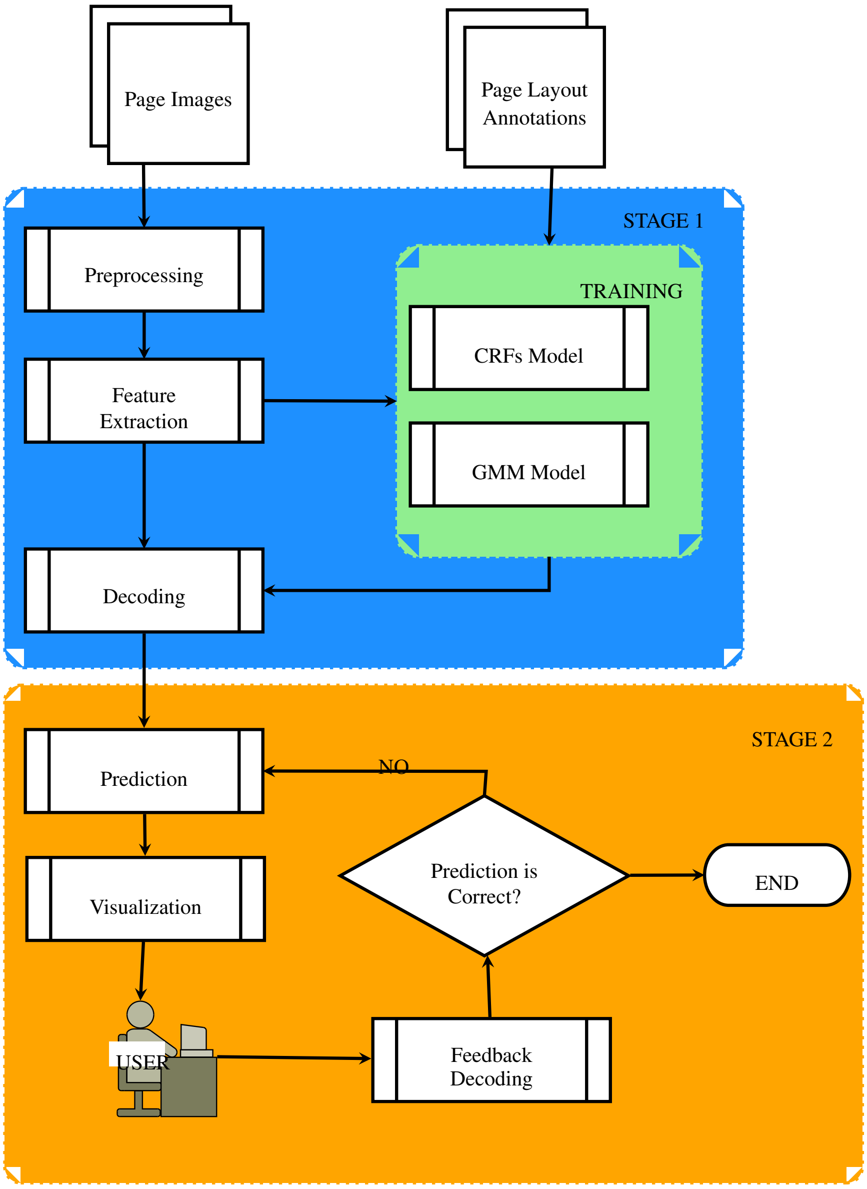


Figure 3.1: System Architecture Diagram

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

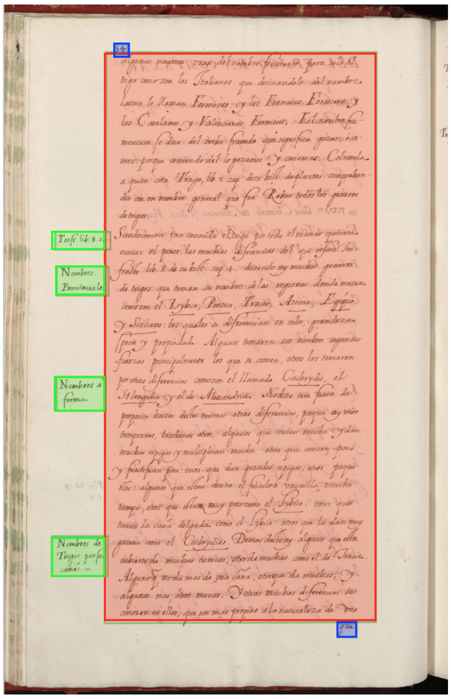
Chapter Outline

4.1 Experiments and Results

4.2 Overview

4.3 Corpus Description

First tome of a seven volume manuscript entitled "Historia de las Plantas" –PLANTAS for short– was selected due it's well structured layout and becouse ground trouht layout is already available. PLANTAS is 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 of PLANTAS consists of a prologue and 152 chapters which make over 1 000 pages with layout ground trouth already labeled using the following categories: catch-word, heading, marginalia, page-number, paragraph, signature-mark, float (drawings); see Figure ?? for reference.



54

algunas naciones vsan del nombre frumento para solo el
trigo como son los Italianos que deriuandole del nombre
latino le llaman Formento: y los Franceses Forment y
los Catalanes y Valencianos Forment. El nombre fru-
mentum ſe dixo del verbo fruendo que significa gozar, ó co-
mer porque comiendo del lo gozamos y comemos. Columela
a quien cita Trago, lib. 2. cap. desu hist. de plantas comprehen-
dio con vn nombre general que fue Robus todos los generos
de trigos.

Teofr. lib. 8. c. — Siendo como es tan conocido el trigo por todo el mundo quisiera
escusar el poner las muchas diferencias del que refiere Teo-
frasto lib. 8. de su hist. cap. 4. diciendo ay muchos generos
Nombres. Prouinciale — de trigos que toman su nombre de las regiones donde nazen.
como son el Lybico, Pontico, Tracio, Assirio, Egipcio,
y Siciliano: los quales se diferencian en color, grandeza.
ſpecie y propiedad. Algunos tomaron sus nombres segun su
fuerzas principalmente los que se comen, otros los tomaron
Nombres á forma. — por otras diferencias como son el llamado Cachrydis, el
Stlengidio y el de Alexandria. No dize sera fuera de
proposito hazer destos mismos otras diferencias, porque ay vnos
tempranos, tardanos otros, algunos que crezen mucho y dan
muchas espigas y multiplican mucho: otros que crezen poco
y fructifican poco: vnos que dan grandes espigas, otros peque-
ñas: algunos que estan dentro el foliculo ó vaynilla mucho
tiempo, otros que estan muy poco como el Lybico: vnos que
tienen la caña delgada como el Lybico: otros que la dan muy
Nombres de Trigos por su cañas. — gruesa como el Cachrydias. Demas desto ay alguno que esta
cubierto de muchas tunicas, otro de muchas como el de Tracia.
Alguno q no da mas de vna caña, otro que da muchas; y
algunos mas, otros menos. Y otras muchas diferencias se
conozen en ellos, que son mas propias á la naturaleza de vno

que

Figure 4.1: Plantas-1 Layout Zones; green for marginalia, red for paragraph and blue for catch-word

4.4 Implementation

4.5 Experiments

4.5.1 Conditional Random Fields

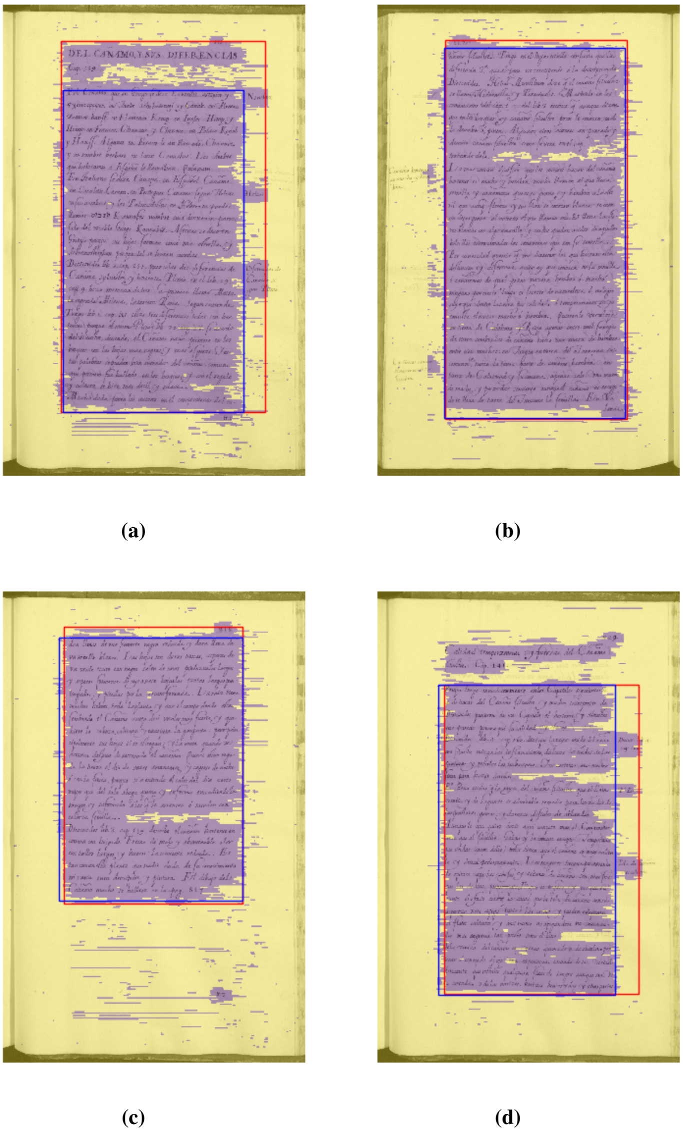
Table 4.1: CRF's Site level Results

Page	Precision	Recall	F1
0944	0.908446	0.914158	0.907493
0945	0.952957	0.952866	0.952905
0946	0.957523	0.955519	0.955996
0947	0.950765	0.951733	0.951163
0948	0.948534	0.950719	0.949605
0956	0.893606	0.896160	0.890686
0957	0.946151	0.945495	0.945381
0958	0.930706	0.927327	0.927419
0959	0.952251	0.951662	0.951921
0960	0.954842	0.953900	0.953925
0961	0.950211	0.950488	0.950346
0962	0.909336	0.920790	0.910282
0963	0.919310	0.914148	0.915994
0964	0.909319	0.910156	0.907617
0965	0.942231	0.941867	0.942029
0966	0.947065	0.946779	0.946860
0967	0.944621	0.946028	0.945210
Global	**0.936154**	**0.936629**	**0.935977**

4.5.2 Site Grouping Approach

In order to have a point of comparison, a simple site grouping rectangle generator have been implemented, this is all adjacent sites classified as "paragraph" are grouped, then we search for the minimum rectangle where all sites of the same group fits and finally we select only the biggest rectangle, see Figure 4.2 for some examples.

Change images to model z0.3_w33_g3



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

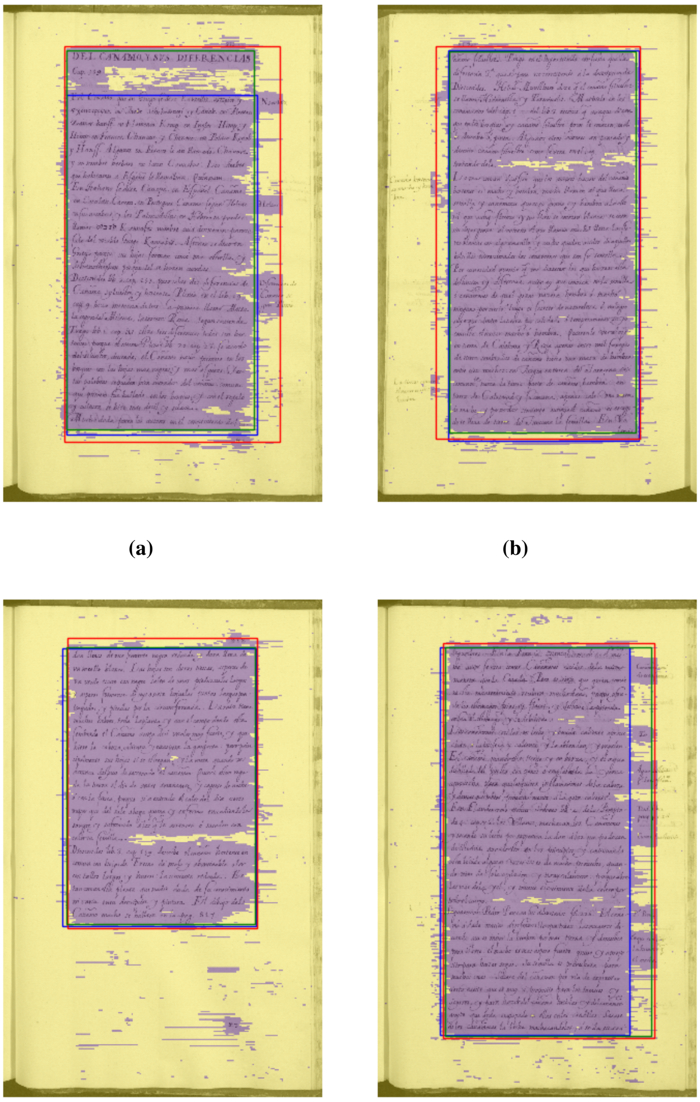
Figure 4.2: Site grouping approach results examples.

4.5.3 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 (Time expended on this stage: $\approx$ 25 seconds using a non-vectorized function). See Figure 4.3

Add numerical results based on [1] method

Remove axis from images



(a) (b)

(c) (d)

Figure 4.3: Proposed method results example, blue line for ground trough, red line for site grouping approach, green line for proposed approach.

4.5.4 Interactive Approach

4.6 Results Discussion

Bibliography

[1] Stamatopoulos, N., Louloudis, G., and Gatos, B. (2015). Goal-Oriented Performance Evaluation Methodology for Page Segmentation Techniques. In *13th International Confrence on Document Analysis and Recognition - ICDAR'15*, pages 281–285.

NOMENCLATURE

CRF	Conditional Random Field
GMM	Gaussian Mixture Model