UNIVERSIDAD POLITÉCNICA DE VALENCIA

Document segmentation using Relative Location Features

Francisco Cruz Fernández

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Supervised by:

Oriol Ramos Terrades
Universidat Autónoma de Barcelona

Enrique Vidal Ruiz
Universidad Politécnica de Valencia

M.Sc. Thesis in Articial Intelligence, Pattern Recognition and Digital Image

Departamento de Sistemas Informáticos y Computación

Valencia, June 2012

Acknowledgements

I would love to dedicate this work to Chloe, for all her support and understanding during this year, and because without her this work could never have been possible.

Also to my parents and to my sister, who are always there when I need them.

Especially thanks to my supervisor Oriol Ramos, for being patient with me and for all the things he has taught me during this year. Also to all the people from the Computer Vision Center for receive me and treat me so well.

Also thank the Universidad Politécnica de Valencia and all the people from the DSIC, especially to Enrique Vidal and all the friends I have there.

Abstract

In this thesis we present a generic layout analysis method devised to work in documents with both Manhattan and non-Mahnattan layouts. We propose to use Relative Location features combined with texture features to encode the relationships between the different class entities. Using these features we build a Conditional Random Field framework that allow us to obtain the best class configuration of an image in terms of energy minimization. The conducted experiments with Manhattan and non-Manhattan layouts prove that using Relative Location Features improves the segmentation results on highly structured documents, as well as results up to the state of the art on documents weakly structured.

Contents

Chapter 1

Introduction

Along the history, humans have felt the need to transmit the information gained through their experiences and beliefs to the rest of people. The emergence of writing abilities in ancient cultures supposed a major breakthrough in the transmission of information, since it allowed written information to be transmitted to many individuals in different places without the need of the presence of the author.

Over the centuries, huge amounts of written documents has been stored, even surpassing the volume of all the current information disposed in digital format. Written documents stored on the world's great libraries and ancient archives offer a high variability respect to the nature of the content and document types. We can find since textual information in the form of scientific studies, literary compositions, administrative documents or records, to maps, paintings, images, photographs or other information provided in document form.

With the objective of preserve all this information, the great breakthrough of scanning technologies in the last years have provided an efficient way to dispose of these documents in digital format while preserving the original source. This situation offers a chance to extract an manage all the information included in these collections, which adds the need of developing methods that help us in this task.

Document Image Analysis and Recognition is the research field in charge of performing the task of analyze and perform overall interpretation of document images. The term of document analysis is a concept that extends to all kinds of documents, since handwritten documents, printed books or forms, to architectural drawings or any other information registered in paper. It is because of the great variability in types of documents that is common to find many works oriented to a particular task, which implies that a method devised to a particular document collection may be initially suitable to work on other documents but probably it will not obtain positive results.

Dealing with documents, one important characteristic to highlight regards to the type of writting. The condition that a document is handwritten or typewritten has several advantages and disadvantages specially regarding to the easiness in the recognition of the text regions or the transcription of the characters. Historical

manuscripts, for instance, may result very difficult to transcribe even for the human eye due to degradation of the pages or complex writting styles. However, this task results less complicated in the case of typewritten documents.

There are several tasks involved in the field of Document Image Analysis which differ in the elements to be analyzed. An important area within this field is text recognition, which is focused in performing text transcription by recognizing the characters and words in documents. Here, we can discriminate between two approaches depending on whether the documents are handwritten or typewritten. In the case of typewritten documents probably the most common technique are Optical Character Recognition (OCR) methods, which perform text transcription by combining word dictionaries with the recognized characters. This methods offer high success rates on typewritten characters, but are not suitable to work on handwritting documents because of the problems described above.

Related with text recognition task, either in printed or handwritten documents, another ongoing research line is the word spotting task. Given an input text string, the objetive of word spotting is to find that inpunt within a colection of documents. Imagine for instance a huge collection of historical documents, the single fact of looking for a particular name in the system and finding the concrete page where this information is shown represent a huge time saving for historians and scientists.

Another important task within Document Image Analysis is the layout analysis, which focus in analyzing the structure of the document by identifying the location of the different elements that comprise it. The layout analysis task usually serves as a previous step for many of the tasks described above, so that the detection of the different entities may suppose a key point in the processing of the document. Unless concrete tasks, general works in layout analysis usually focus on the detection of two main classes of elements, the ones belonging to text regions and the ones containing graphical elements [17].

1.1 Related Work

The problem of layout analysis have been addressed in several ways along the last decades. One important issue to take into account in the election of the method is the type of documents for which the method is devised. Within the field of layout analysis, we can mainly consider two groups of document layouts according to the distribution of the elements that comprise them: Manhattan and non-Manhattan layouts.

Documents with a Manhattan layout are often organized following a grid shape or another regular structure, some cases of journal documents with rectangular text regions can be included in this category. On the contrary we can find the documents with a non-Manhattan layout, which are identified for being composed by irregular regions and without a visible structure.

From the possible ways to address layout analysis on both types of documents we can mainly distinguish between three families of methods: connected compo-

nents, projection-based and texture analysis.

On the one hand, methods based in connected components have proved to offer great applicability on both Manhattan and non-Manhattan layouts. This methods usually follows a bottom-up approach where pixels are grouped into more complex structures until the final entities are identified. An example in this line is the work of Li *et al.* in [23], where connected components are extracted and clustered into a tree description to perform page segmentation in journal documents. A similar approach is presented in the Océ method [4] where the authors make use of a decision tree classifier based in connected components to detect text regions, images and other elements in contemporary documents.

On the other hand we can find methods based in projection profiles. In [6] a method for text line detection in historical documents is presented. The authors propose a model inspired in natural language processing methods combining both vertical projection profiles and Hidden Markov Models. Also in [11], Ha *et al.* present a method based in recursive X-Y cuts for layout analysis in journal documents. This method is based in decomposing the document image into a set of rectangular blocks and performing vertical or horizontal projection profiles for the estimation of the layout. The work of Nagy *et al.* in [16] is another example in this line, in this work they also use the X-Y cut approach to identify text regions in technical reports. Unlike methods based in connected components that can be applied on several layouts, this approach does not produce good results on documents with a non-Manhattan layout. However in the case of Manhattan layouts results to be very useful.

In addition to these methods, the third very extended family of methods on document segmentation is the analysis of the image texture. These methods are characterized for offering great invariance to possible changes in the possition and shape of the elements to detect, and can be applied on both Manhattan and non-Manhattan layouts with good results. One work that uses these features in text segmentation is [14], where Jain *et al.* perform text segmentation from journal pages using Gabor filters as mechanism of texture analysis. Another work in this scope is [17], which also uses texture information via Matched Wavelets to obtain text segmentation in both real scene and journal documents. The great potential shown by texture-based methods in image segmentation and object detection, [15], has promote several works comparing different texture methods [21]. Here the authors conclude that there is not a favourite method for every task, but each of them offers partitular advantages and disadvantages according to the image texure that have to be considered.

Most recently, it have been proved that the inclusion of contextual information helps to improve the obtained results in most pattern recognition tasks, and layout analysis in documents is not an exception. Many works have tried to include this additional information into their models in form of neighbour relationships, and one common way of doing that is modeling the pixel dependences using Conditional Random Fields (CRF). Conditional Random Fields belongs to the family of graphical models that permits to model complex dependences between the random

variables of the problem. Introduced by Lafferty *et al.* in [19] in the context of labelling sequence data, CRFs have been stablished as a powerfull method in image segmentation in the modeling of these dependences. There are several works in the field of document segmentation that use CRF to refine their results by including pairwise dependences between neighbor pixels [17, 20], further but related to document segmentation, it have been also used in object segmentation tasks [9].

In addition to modeling pairwise dependences, other works try to model the spatial relationships between classes as way to indicate the location where it is most likely to find elements of a particular class. This approach was intially presented in [10] where the authors present Relative Location Features as a mechanism to encode this information. To encode this features they compute probability maps as a way to represent how likely is to find an element of one class in a particular position regarding to the rest of the classes and then include this preferences into the final features.

We have seen that a lot of different approaches have been developed in the field of layout analysis in document images, however these methods are usually devised to work on a specific type of documents or dataset, so there is not a direct way to check if one particular methods works generally better than other. In the last years several contest in the field of document segmentation have been celebrated providing a way to face this situation [4, 2, 3].

1.2 Objectives

In this thesis we address the problem of layout analysis on historical documents. We want to identify and extract the different entities that make up the documents on which we will work. To perform this task, we propose a texture-based segmentation method and the utilization of the Relative Location Features into a Conditional Random Field framework. In addition, our method is devised to offer great applicability on different layouts, for what we will use a benchmark dataset to prove its effectiveness.

The objectives of this thesis, therefore, can be summarized as:

- Perform layout analysis in historical documents using a Conditional Random field framework.
- Check whether the inclusion of Relative Location Features helps to improve the segmentation results either in structured historical documents as in a contemporary benchmark dataset.
- Compare the effectiveness of our method on documents with a manhattan and non-manhattan structure.

1.3 Structure

The rest of this thesis is structured as follows: Chapter 2 describes the provided datasets for the realization of this thesis. Chapter 3 describes the theoretical framework of our method, and describes the used techniques as well as the construction of the Relative Location Features and its inclusion into the CRF framework. The proposed experiments and a discussion about the obtained results are shown in chapter 4. Finally, conclusions about this work and the open research lines are commented in chapter 5.

Chapter 2

Datasets

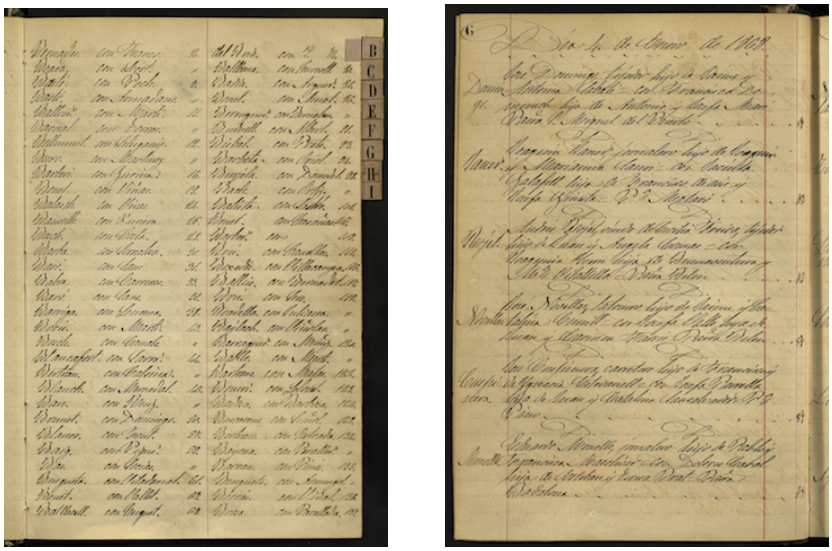
In this chapter we describe the different datasets used in the realization of this thesis. We wanted to test our method either on documents with a Manhattan and non-Manhattan layout, so we have chosen two different sets of documents that ajust to this condition. First, we present the 5CofM project which have motivated the realization of this work. The documents on this dataset show a Manhattan layout and results very appropiate to test the effectiveness of the Relative Location features. Second, we descibe the PRImA Layout Analysis Dataset, which can be considered as a benchmark dataset that will permit us to objectively compare our results with the state of the art methods. The documents on this dataset corresponds with a non-Manhatan layout, and will be useful to check whether our method is suitable for general layout analysis tasks.

2.1 Five Centuries of Marriages project

Five Centuries of Marriages (5CofM) is a long-term research initiative based on the data-mining of the *Llibres d'esposalles* conserved at the archive of the Barcelona Cathedral. This collection is composed by a set of 291 handwritten books including marriage licences conducted in the period from the year 1451 until 1905. The set of books includes approximately 550,000 marriage licences from 250 different parishes in the Barcelona area. The entire collection was digitalized and stored in color images represented in PNG format without compression with a resolution of 300 DPI, which ensures enough image quality for analyzing and processing all the contents adequately.

The information included on this collection has a huge value for the conduction of social and demographic studies. From the study of this information is possible to learn about migration flows, socioeconomic position of the families, geographic distribution or genealogical studies among others. All this information is suposed to be part of a exhaustive database named *The Barcelona historical Marriage Database*.

The project is conducted by the colaboration between the Center for Demo-



(a) Index page (b) Licences page

Figure 2.1: Examples from both types of pages. (a) shows a page from the index from volume 208. (b) shows a license page from the same volume

graphic Studies (CED) and the Computer Vision Center (CVC) from the Universidad Autonoma de Barcelona (UAB), and suppose the establishment of research links between the areas of social science and computer science.

Each book of the collection is composed of two different types of pages. The first one correspond with the index of the licenses contained in the book. Each entry of the index is formed by the husband's surname followed by the number of the page where the license is located, besides, depending on the period when the book was written is possible to find also the city where the marriage was conducted. The entries are organized in a set of columns along the page separated by a small blank space (see Fig 2.1a).

The second type of page includes the marriage licenses. Each page of the set is formed by a variable number of licenses arranged along the page (see Figure 2.1b). Each of them is composed of three elements that identify and virtually divide the license into three columns. From left to right we can see, firstly, the husband surname, secondly, the license body, that includes the name of the grooms and their parents, husband's occupation, place of the marriage and other variable information, and thirdly the paid tax for the marriage.

Even though the final objetive of the 5CofM project is to process both types of pages, in the present work we focus on the pages containing the marriage licenses. Thus, in the case of these pages, we consider a total of 4 different classes for the segmentation process: *name, body, tax and background* (see Figure 2.2).

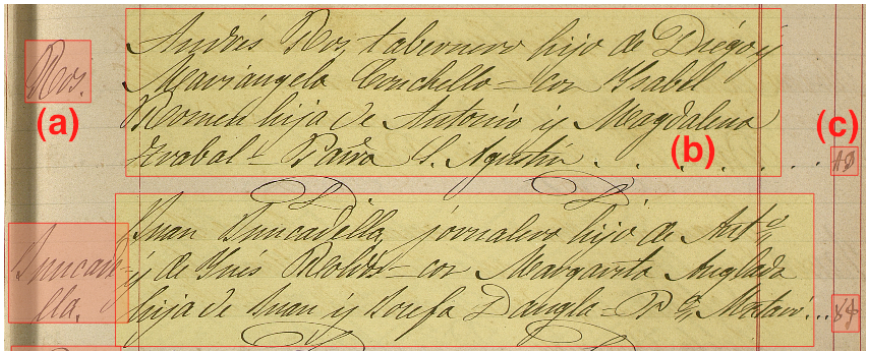


Figure 2.2: Two records from 5CofM database volume 208 showing the structure of the classes: (a) Name (b) Body (c) Tax

2.2 5CofM Ground Truth tool

Now we describe the construction process of the ground truth especially developed for this collection. As we described before, 5CofM is a recent ongoing project, so there is no ground truth available and adjusted to our needs. As a part of this thesis, we labeled a set of pages which will be used to perform the experiments of this work and future ones.

To carry out the labelling process of the provided set of images, from the Computer Vision Center (CVC) it has been developed a web-based crowsourcing tool that allows to perform the labeling of the elements of interest for the different tasks that are going to be performed over this collection. The tool is devised to receive the contributions from any person interested in participate in the project, either by tagging the different elements or making text transcription.

In this work, we have manually tagged a total of 80 pages from the volume 208 of the collection. The labelling process of the elements that we nedded for this task consist in marking each of the three regions of interest on each license by drawing the corresponding bounding boxes as is shown in Figure 2.2. Thus, the tool registers the bounding boxes drawed over each license for each page of the collection. A screenshot from the labelling process of one page can be seen in Figure 2.3.

Once we have tagged all the licenses of one page, the application provides of a SVG file with all the registered information of the page. This output file contains the coordinates of the upper corners from the different bounding boxes and its vertical and horizontal length. The hierarchical distribution from the SVG tags allow to arrange the file in a set of licenses, which results very useful for our task.

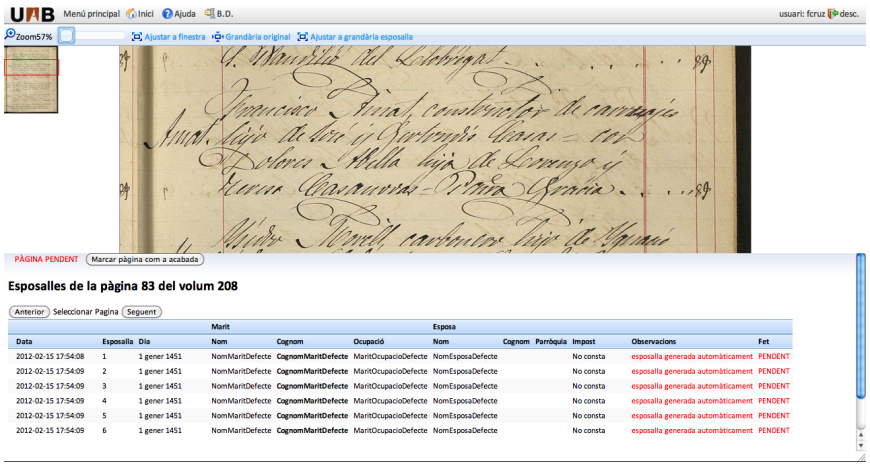


Figure 2.3: GroundTruth tool

2.3 PRImA Dataset

The second dataset we present correspond to the PRImA dataset. This dataset was developed by a research team from the *University of Salford* with the objective of provide a realistic dataset on which different layout analysis methods can be evaluated under the same conditions. This database was presented in the 10th International Conference on Document Analysis and Recognition (ICDAR) [1] and has been stablished as a reference corpus in the layout analysis field also present in the ICDAR contest as a benchmark dataset.

The dataset currently includes 1240 color images from several types of printed documents where it is possible to find since magazine pages to technical articles, bank statements, forms and advertisements. The images are provided in TIF and JPG format with a resolution of 300 DPI. The great variety on kinds of documents makes this corpus to treat with some challenging situations in layout analysis, as it include for example many instances of text in regular and irregular text boxes with different font sizes, images including text, or graphics. Two examples of images from the dataset can be seen in figure 2.4.

We choose this dataset in this work for several reasons, on the one hand the test partition that we have to our disposal is the same used in the ICDAR contest, which means that our results could be directly compared to the results on previous contest. On the other hand, in contrast to the 5CofM documents where it is possible to appreciate a Manhattan layout, images in PRImA dataset are different from each other and there is not a defined pattern on each of them. This condition will permit us to evaluate as far as the Relative Location Features helps to improve the segmentation results on Manhattan and non-Manhattan documents. For this dataset, we consider 3 different classes in the segmentation process: *text, image and background*.

Figure 2.4: Four examples of images from the PRImA dataset

Chapter 3

Method

In this chapter we describe the developed method for this our document segmentation task. The objective that we follow is to find the optimal labelling configuration for an image I that maximizes the *a posteriori* probability $P(C|I)$, being C the set composed of the label values corresponding to each pixel p in the image, considering the set of l class labels $L = \{c_1, c_2, ..., c_l\}$. The proposed method aims to compute this probability by means of a Conditional Random Field framework. Using this model we are able to encode the pairwise dependences between the image pixels and compute the previous probability distribution in terms of energy minimization following the expression:

$$P(C|I) = \frac{1}{Z} \exp \left\{ - \sum_i D_i(c_i) - \sum_{\{i,j\} \in N} V_{i,j}(c_i, c_j) \right\} \tag{3.1}$$

Let analyze each term in detail. Firstly, the unary potential $D_j(c_j)$. This term represents how well the label c_j fits in the j-th pixel p_j of the image I. Secondly, we have to model the pairwise pixel interaction modeled by the energy term $V_{i,j}$, which represents the pairwise potential between the set N of interacting neighbor pixels. This potential represents the penalty of assigning pixel label c_i and c_j to the pixels p_i and p_j respectively. We decided to model these penalties by means of the Potts model as $V(\alpha, \beta) = T_{\alpha,\beta} \cdot \pi(\alpha \neq \beta)$ where the factor $\pi(\cdot)$ is 1 if the condition is true and 0 if false, and $\alpha, \beta \in L$. The use of this function must favors the pixels included on the same regions to be labelled with the same label, while obtaining better defined boundaries between regions. Finally Z is a normalization defined in next sections.

In addition, note that the model was initially proposed to work at pixel level, however, as we deal with large images is necessary to make some modifications to ensure that the segmentation process is efficient. Our choice to overcome this situation was based in consider groups of pixels rather than individual pixels. It was decided to use a regular grid system that divide the image into regions of regular size called cells. Thus, we defined the grid $S_k \in \{S_1, ..., S_K\}$ consisting

of K rectangular cells covering the entire image. Given this new distribution of the image, the goal of finding the optimal pixel labeling is replaced by the goal of finding the optimal labeling at cell level. It is understood that this process may undermine the accuracy of the segmentation process, however we believe that does not substantially affect the objectives of this thesis.

The rest of the section if organized as follows: Section 3.1 describes the Conditional Random Field theoretical framework whitin the graphical models family. Then, the main characteristics of the chosen minimization algorithm are described in section 3.2. Finally, the proposed Gabor features and the Relative Location Features are described in sections 3.4 and 3.5.

3.1 Conditional Random Fields and graphical models

Many problems in the field of the computer vision require to model the existing dependences between the set of random variables $X \cup Y$, considering X as the set of real-valued feature variables and Y the set of class variables to predict. Thus, in our segmentation task the label assigned to a group of pixels is supose to affect to the decision of labelling the pixels in their neigborhood by being related with each other. Therefore, the objective is to model the probability distribution of Y in respect to the variables on X taking into account this dependence level between them.

When we face with this situation is neccesary to dispose of a method that permits to model this dependences, however this task may become complicated in cases with a large amount of variables or complex dependences between them. Graphical models provide a consistent framework in the modeling of the probability distributions behind these dependences. Acording to the dependence level that we want to model, there exists a entire family of graphical models that adjust to each of the possible circumstances.

One of the most basic models is the naive Bayes model, where all the input variables are independent from each other. This model provides the neccesary to model the joint distribution between the feature variables of the model an its correspondent class variables. Extending this model to the next level of dependence, we can consider the case where the variables are arranged in a sequence, and here the Hidden Markov Models (HMM) are one of the most appropiate models to deal with it. This model is widely used in speech and handwritting recognition problems by its ability of model the temporal relationships between the elements of the speech or text sequences. The next logical extension for the dependences between the variables can be represented as relationships between the nodes in a graph. Here, we can locate several methods which differ in the topology of the graph edges and variables, as Bayesian networks, Markov Random Fields or Conditional Random Fields.

Despite the similarities in the graphical representation, there are some differences between these methods in most part due to the probability distribution that

model and the type of dependences between the variables. The main idea under graphical models to model this distribution is that a probability distribution often can be factorized in a set of local functions that depend of a smaller subset of variables, which we call factors. Thus, considering the set of variables $X \cup Y$, the probability distribution among them can be formulated as a product of factors $\Psi_j(a, b)$ in the scope $a \subseteq X \cup Y$ as:

$$p(Y, X) = \frac{1}{Z} \prod_{j \in \varepsilon} \Psi_j(a, b) \tag{3.2}$$

where ε represents the set of dependences between a, b and the constant Z, called partition function, is a normalization factor to ensure that the probability distribution sums 1 defined as:

$$Z = \sum_{x,y} \prod_{j \in \varepsilon} \Psi_j(a, b) \tag{3.3}$$

we can realize as the calculation of Z is based on all the possible values of X and Y, which correspond with a exponential number of combinations between all possible the values of those variables. In the computation of this value lies the main complication in the inference of these models.

In problems where we deal with images the set of input variables X is already known, so we can keep it fixed. The resulting model is a Conditional Random Field defined from Eq. (3.2) as:

$$p(Y|X) = \frac{1}{Z(X)} \prod_{j \in \epsilon} \Psi_j(a, b) \tag{3.4}$$

where the partition function is now defined considering all the possible configurations of the class variables Y as:

$$Z(X) = \sum_{Y} \prod_{j \in \epsilon} \Psi_j(a, b). \tag{3.5}$$

Thus, this partition function correspond with the normalization factor defined in Eq. (3.1) according to the particular set of features X from each image cell.

Initially we can consider the set of functions Ψ_j as composed by any kind of function, however the use of exponential functions offers several advantages in our task. First, the use of exponential functions allow us to define our model in terms of summations instead of products. Second, it is easy to represent them as energy potentials.

3.1.1 Inference in graphical models

Efficient inference of the model parameters may be a difficult process according to the dependence level between variables. Inference on sequence-based models as Hidden Markov Models or linear-chain CRFs have proved to be exact using

Belief Propagation based algorithms, as well as for the estimation of the most likely assigment class variables by means of the Viterbi algorithm.

In the case of general graphical models, computing exact inference results to be intractable in the case of having loops. For these situations, a family of inference methods have been developed to offer an approximate estimation of the model parameters. Loopy Belief Propagation algorithm is the generalization of the general Belief Propagation for these models.

The estimation of the most likely assigment for the model variables can be also computed by means of approximate methods as the max-product algorithm. In addition, the use of log-linear functions allows to easily define the maximum a posteriori estimation of Eq. (3.4) in terms of energy minimization as shown in Eq. (3.1). Here, methods as the Graph Cut algorithm can be used for the estimation process of this energies.

3.2 Energy Minimization via Graph Cuts

We have proposed a CRF framework to perform layout analysis in terms of energy minimization as stated in Eq. (3.1). We have seen that a exact estimation of the best labeling can not be computed efficiently due to the great number of variables in the model, so we were forced to search for alternative approximate methods.

Graph Cut algorithm [7] allow us to efficiently compute a local minimum of the minimization function based in the minimum cut problem of graphs theory. The idea behind this method lies in represent both the feature and class variables within a graph structure and define the most likely assigment as a cut in the graph.

In this way, the CRF is represented as a undirected weighted graph $G = \{V, \varepsilon\}$ with two different types of nodes. First, we find the nodes that correspond to each input variables of the model, in our case each image cell S_k. Secondly, the graph includes the nodes correponding to the set of possible labels L to be assigned to each variable, called *terminal nodes*. Each node will be connected by means of two types of edges. It is called *t-link* to an edge that connects a terminal node with a cell node, on the same way we call *n-link* to the edge connecting two neighbor cells nodes according to chosen connectivity. Finally, we define C as the set of labels assigned to each cell at the end of the process, so that C_k will represent the assigned label to cell k. A visual representation of the structure of the graph can be seen in Figure 3.1.

Formally, given two terminal nodes c_1 and c_2 and the set of cells S, we call S_1 to the cells (nodes) assigned with the assigned label c_1, and in the same way, S_2 to the nodes with labeled with c_2. Thus, the set of nodes $V_{1,2}$ is formed by the terminal nodes c_1,c_2 together with $S_{1,2} = S_1 \cup S_2$. In what regards to the edges, the term t_i^1 represent the *t-link* between the node i and the terminal node c_1, in the same way we call t_i^2 to the *t-link* joined to the terminal node c_2. We also call $e_{\{i,j\}}$ to the *n-link* connecting two neighbor pixels i, j.

Initially, the graph $G = \{V, \varepsilon\}$ includes all the *n-link* according to a 4-connectivity

and all the possible *t-link*. So, the goal of the algorithm is to find a cut $\zeta \subset \varepsilon$ of minimum cost resulting in the induced graph $G = \{V, \varepsilon - \zeta\}$. The cost of a partition is given by the sum of the wheights of each edge included in the resulting graph, this wheights are defined as follows:

- $t_i^1 = D_i(c_1) + \sum_{j \in N_i} V(c_1, c_j)$
- $e_{\{i,j\}} = V(c_1, c_2)$

In that way, the total cost of C is given by the following expression:

$$E(C) = \sum_{i \in S} D_i(c_i) + \sum_{\{i,j\} \in N} V_{i,j}(c_i, c_j) \tag{3.6}$$

which corresponds with the argument of our exponential function defined in Eq. (3.1).

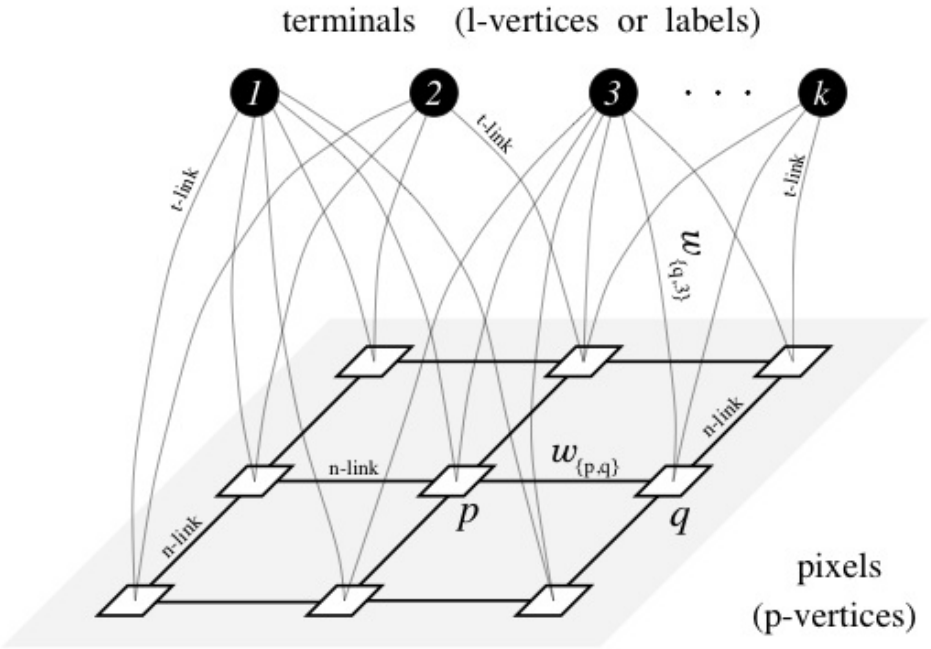


Figure 3.1: Representation of the graph $G = \{V, \varepsilon\}$ from the GC algorithm [7]

3.3 Proposed method

We propose two different approaches to compute the unary term $D_i(c_i)$ in Eq. (3.1) to evaluate the contribution of the Relative Location Features.

Our first approach is based in use texture features to compute a cell-level classification, using the conficence of the classifier for the construction of the CRF. Then, in our second approach we include the Relative Location Features into the final model. Figure 3.2 shows a diagram with the proposed approach using the RLF.

The following sections describe in detail each of the proposed approaches.

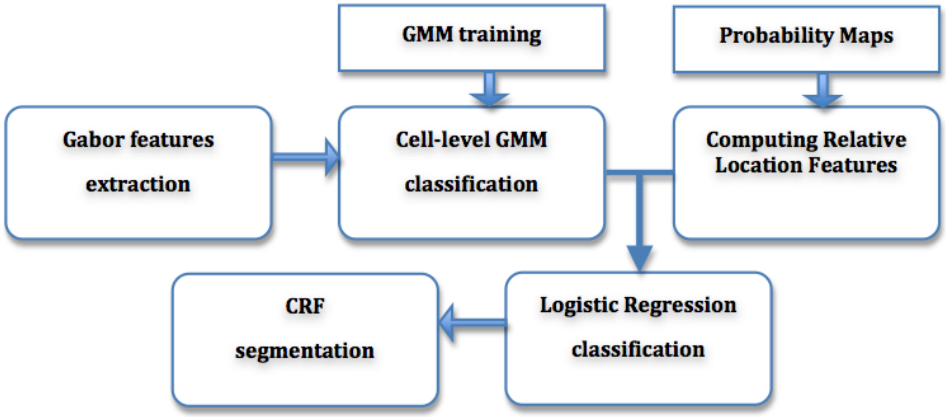


Figure 3.2: Segmentation process using Relative Location Features

3.4 Gabor features

Gabor filters, named by the physicist Dennis Gabor, are linear filters composed by a sinusoidal wave modulated by a gaussian envelope that allows to perform measurements in both the spatial and frequency domains, which is of great worth for a segmentation task. It have been proved that some impulses of the visual process in mammalian brains can be modeled by Gabor functions [8], and the this method shares similarities with how humans process the information when analyze and discriminate between the different elements in an image.

Gabor filter was initially proposed to work on 1D signals, but in our case we will work with the expanded 2D model. 2D Gabor filters are composed by the product of an elliptical Gaussian and a complex exponential representing a sinusoidal plane wave. The filter in the spatial domain is represented as follows:

$$\psi(x, y : f_0, \theta) = \frac{f_0^2}{\pi\gamma\eta} e^{-\frac{f_0^2}{\gamma^2}x'^2 + \frac{f_0^2}{\eta^2}y'^2} e^{j2\pi f_0 x'}$$
$$x' = xcos\theta + ysin\theta$$
$$y' = -xsin\theta + ycos\theta$$

where f_0 de filter base frequency, γ and η correspond with the perpendicular axes of the gaussian θ. In the frequency domain, the Fourier transformed version of the filter correspond with a single Gaussian bandpass filter defined as:

$$\psi(u, v|f_0, \theta) = e^{-\pi^2(\frac{u'-f_0}{\alpha^2} + \frac{v'}{\beta^2})}$$
$$u' = ucos\theta + vsin\theta$$
$$v' = -usin\theta + vcos\theta$$

Usually, when we are using filter responses as features in a classification problem, the common procedure is computing a set of responses by means of a Multi-Resolution Gabor filter Bank. A filter Bank contains a set of filters implemented with different orientations and frequencies, varying this parameters ensures the detection method to be invariant to position or orientation changes in the image elements.

Some previous works have studied the best way to select the set of orientations and frequencies [18]. Following their indications, we consider a exponential spacing of m frequencies $f_l = k^{-l} f_{max}$, with $l = \{0, ..., m-1\}$, where f_{max} is the maximum frequency fixed, f_l the lth frequency and k the scaling factor $(k > 1)$. Regarding to the orientation angle, the most common is to set orientations uniformly as $\theta_b = b2\pi/n$, with $b = \{0, ..., n-1\}$. Where n is the number of orientations to be considered.

In this way, we define our feature vector as the set $x \in \mathbb{R}^{m \times n}$, where the feature $x_{l,b}$ is given by the filter response of frequency l and orientation b on a particular image pixel.

Although the advantages of Gabor filters, there are some considerations to take into account. The main complication in the use of these filters is the computational cost of computing the responses for all the filters in the bank. The involved Fourier transform, for example, is one of the most expensive process to be computed in the computation of a Gabor response. In this field, the work of J. Ilonen *et al.* will help us to obtain this responses in a efficient way [13]. We use the Multi-Resolution Gabor Filter toolbox developed in [13] for the realization of the present work.

Recent studies suggest that analyzing the relationships between filter responses helps to discriminate between different objects or classes in a classification task [12], and here, statistic models are the most appropiate to capture this information in detail. Among the possible models, Gaussian mixture models appears to be a powerfull method to capture this information.

3.4.1 Gaussian mixture model

Once we have established how we compute the feature vectors, we need to decide which method is more appropiate to estimate a proper probablity density function (pdf) over the training data. Although there are many methods that could be used to estimate this pdf, it have been proved that finite mixture models approximate more complex, and therefore, more precise pdfs [22]. Here, among the possible distribution functions, Gaussian mixture models (Gmm) are widely used in these tasks, and have proved to be a powerfull tool when we work with continuous-valued feature vectors.

A gaussian mixture model is defined as the result of a weighted sum of M gaussian components as:

$$p(x|\theta) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{3.7}$$

where x is the feature vector described above and the factor w_i corresponds with the weight that the component i have over the feature vector x. These weights are restricted to satisfy $\sum_{i=1}^{M} w_i = 1$. Each density function $g(x|\mu_i, \Sigma_i)$ correspond to one of the M gaussian components given by the expresion:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \tag{3.8}$$

where μ_i and Σ_i are the mean and covariance matrix of the component i, repectively.

The full Gaussian Mixture model is finally defined by the paremeter list $\theta = \{w_1, \mu_1, \Sigma_1, ..., w_M, \mu_M, \Sigma_M\}$ that have to be estimated from the training data. The estimation of these parameters is performed by the Expectation-Maximization (EM) algorithm.

3.4.2 Gabor-based classifier

Now we show how to include the obtained information by the Gmm into the proposed CRF framework. In this model we model each image pixel by its feature vector x_j as described before, in that way, the probability $P(c_j|p_j)$ in Eq. (3.1) that we compute is now defined as $P(c_j|x_j)$, which is computed through the Gaussian Mixture model.

Previously we described that we changed the image representation from pixels to a set of cells, so, as the described Gaussian Mixture Model provide us with a pixel level classification, is necessary to define a way to obtain this information but focused on each image cell S_k. In this scope we have studied three different approaches for the computation of $P(c_k|S_k)$:

The first approach was devised to take into account the pixel level classification given by the probability density function from the Gaussian Mixture Model. One

we have a label assigned to each pixel of the cell, we assign the cell label as the mayority class between the pixel labels as:

$$P(c_k|S_k) = \frac{\#\{p \in S_k | c_k = \arg\max_c p(x_k|c)\}}{\#S_k} \tag{3.9}$$

Secondly, we decided to take into account the prior probability of the occurrence of each class along the different document pages. Thus, the probability is now defined as:

$$P(c_k|S_k) = \frac{\sum_{j \in S_k}(P(x_j|c_k))P(c_k)}{\sum_c \sum_{j \in S_k} P(x_j|c)P(c)} \tag{3.10}$$

One problem from the previous approach is that some classes have very small prior probabilities. This event can be explained just by taking a look to the document images. We can see in Figure 2.2 as the classes *tax* and *name* include a reduced number of pixels respect to the class *body* or the background, therefore, as the prior probability is computed by the proportion of pixels of each class, the computed value results to be very small. To avoid this problem, we consider as third approach that all classes are equally probable, so, the proposed model results in:

$$P(c_k|S_k) = \frac{\sum_{j \in S_k}(P(x_j|c_k))}{\sum_c \sum_{j \in S_k} P(x_j|c)} \tag{3.11}$$

Once we have computed the value of $P(c_k|S_k)$ through any of the previous approaches, the unary term $D_k(c_k)$ in Eq. (3.1) is computed as $D_k(c_k) = -\log P(c_k|S_k)$.

3.5 Relative Location features

So far, we have seen that using Conditional Random Fields on a segmentation task allow us to encode local and pairwise preferences between the model variables. However, this relationships are often limited to the variables in a defined neighborhood system, but although this information results to be useful in the segmentation process, there are still another relationships that can be model and to enrich the final model.

There are some tasks where a previous knowledge about the semantics of the classes can be used to guide the segmentation process, for instance, if we want to discriminate between the classes *body* and *tax* (see Figure 2.2), we can restrict the system to find *tax* pixels only in areas above and at the right side from *body* pixels. There are also some cases where we do not have a semantic knowledge but we know from the definition of the problem that there is a correspondence between the classes. In administrative documents, for instance, we may know that the signature appears always in some place at the bottom of the document.

Following this approach, Gould *et al.* present in [10] a method that permits to capture the inter-class spatial relationships and encode it in a feature. The method

is devised for a multi-class segmentation problem in real scene images where there exists a previous knowledge about the most common location and the semantic of the classes.

For example, some images in this work contains pixels from the classes *tree, grass, sky*, so we can sense that *sky* pixels will appear, in general, above pixels of the class *grass*, as well as pixels from class *grass* will appear around the class *tree* and below *sky*. All this constraints are encoded by the proposed Relative Location Features.

The implementation of this features includes several steps. Firstly, we compute a probability map for each pair of classes, which encodes the probability of finding items of one class in a particular location according to the relative position of elements of another class. Secondly, we need an initial estimation of the each cell class, which can be usually performed by means of a appearance-based classifier, although any other classifier can be used. The meaning of this step is to have a starting point where to apply the knowledge about the spatial relationships, and then correct the inicial labeling according to it. Finally, combining the information provided by the initial prediction with the probability maps we will encode the Relative Location Features. Each of these steps are described in detail below.

3.5.1 Probability maps estimation

Probability maps are a mechanism that allow us to represent the location where is most probably to find elements of one class respect to the location of an element of another class. In other words, given two classes c_i and c_j, and one pixel from each class p_i and p_j, we say that the map $M_{c_i|c_j}(u,v)$ encodes the probability that a pixel p_i, with a relative displacement (u,v) from p_j, belongs to the class c_i as:

$$M_{c_i|c_j}(u,v) = P(c_i|p_i,p_j,c_j) \tag{3.12}$$

We need to learn a probability map for each pair of classes, including the map of one class respect to itself. Besides, the maps are not reciprocal, *i.e.* $M_{c_i|c_j}(u,v) \neq M_{c_j|c_i}(u,v)$, so that the neccesary number of maps to compute will be, considering l different classes, l^2 maps.

The learning process of the probability maps is carried out by counting the number of times that each possible displacement between pixels from the two different classes is given over the training set. More concretely, the map $M_{c_i|c_j}$ will be computed by calculating all the displacements between the pixels from the class c_i and the pixels from the class c_j. For example, let the pixels $p_i = (x_i, y_i)$ y $p_j = (x_j, y_j)$, the displacement between them is $(u,v) = (x_i - x_j, y_i - y_j)$. Once we have calculated the previous displacement, the value of the map $M_{c_i|c_j}$ at the position (u,v) is increased by one. Repeating the same calculation for each pair of pixels from the classes c_i and c_j we will obtain a region of the map where the majority of the votes are grouped, which correspond to the most probably area.

Computing all the possible displacements between pixels of one class (original)

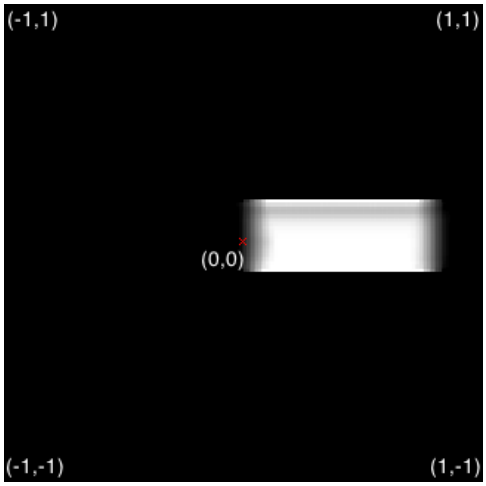


Figure 3.3: Result of computing the Map $M_{body|name}$ on the 5CofM licenses

with respect to another (reference) may result computationally expensive. In order to reduce the number of calculations we substitute the pixels from the reference class by the centroids of the cells that comprise it. As we still consider all the pixels of the original class, this process does not substantially affect to the global result of the map.

3.5.2 Encoding Relative Location Features

At this point we have computed the Probability Maps for each pair of classes, and we have a method to perform a first class prediction for each image cell, so that we can proceed with the encoding of the RLF.

We start from a image representation consisting in a set of cells S_k, and we have calculated a initial class prediction $\hat{c}_k$ for each of them with a probability $P(\hat{c}_k|S_k)$ given by our Gabor classifier. Now, the idea of the RLF is to provide a set of features to each cell which model the probability of belonging to each of the possible classes taking into account the class labels from the rest of the cell images. Therefore, considering our l different labels, at the end of the process we should have l Relative Location Features for each image cell.

The construction process of RLF is conducted as follows: each image cell will cast a total of l votes over each of the other cells. These votes can be understood as how likely is to assign a particular label to a cell, according to the initial labelling assigned to them and the information provided by the probability maps. Thus, if we have K different cells, each cell will recibe $K - 1$ votes from the rest of them.

In order to get more profit from the RLF and following the approach in [10], we decide to split the process to compute two different RLF called v_{other} and v_{self}. The first one gather the votes from cells with different initial class label, and the second one the votes that come from cells with equal labels. That division will

allow us to assign different weights to each feature in the incorporation to the Conditional Random Field. The construction of these features is formulated as:

$$\begin{aligned} v_c^{other}(S_i) &= \sum_{j \neq i : c_i \neq \hat{c}_j} \alpha_j M_{c_i|\hat{c}_j}(x_i - x_j, y_i - y_j) \\ v_c^{self}(S_i) &= \sum_{j \neq i : c_i = \hat{c}_j} \alpha_j M_{c_i|\hat{c}_j}(x_i - x_j, y_i - y_j) \end{aligned} \tag{3.13}$$

where (x_i, y_i) and (x_j, y_j) are the coordinates from the centroids of the cells S_i and S_j, respectively. Votes are also weighted by $\alpha_j = P(\hat{c}_j|S_j)$, which is provided by one of the equations shown in Section 3.4.2. We can appreciate with the inclusion of the α_j factor the importance of providing a confident initial class prediction.

As a last stage, we add a normalization step to define a proper probability distribution over the new features. In this case we normalize to ensure that $\sum_{c=1}^{l} v_{c_k}^{other}(S_k) = 1$, and respectively for the values of $v_{c_k}^{self}$.

3.6 Combination of features

Now we show how to include the Relative Location Feature previously computed in Eq. (3.13) into the CRF model described in Eq. (3.1). In this case the unary term $D_k(c_k)$ will include all the information gathered at this point of the segmentation process, this includes both the new Relative Location Features and the initial class prediction based on the Gaussian Mixture Model. More concretely, Relative Location Features are linearly combined with the appearance-based features as follows:

$$\begin{aligned} D_k(c_k) = w^{app} \log P(c_k|S_k) + \\ + w_{c_k}^{other} \log v_{c_k}^{other}(S_k) + w_{c_k}^{self} \log v_{c_k}^{self}(S_k), \end{aligned} \tag{3.14}$$

where the weights w^{app}, $w_{c_i}^{other}$ and $w_{c_i}^{self}$ are learned using a logistic regression model from he training dataset.

Chapter 4

Experiments and results

This chapter describes the conducted experiments to evaluate our proposed method. In order to evaluate the use of the RLF in a document layout analysis task we will apply our method on two different datasets that show a Manhattan and non-Manhattan layouts. As a example of Manhattan layout we used the 5CofM dataset presented in Section 2.1, and the PRImA dataset as a case of non-Manhattan layout as well as a benchmark dataset to compare our method with other works in document layout analysis [5].

First, we explain the metrics used to evaluate the quality of our results, then we describe the training and test partitions on each of the datasets. Next, we show the experimental settings of the Gaussian Mixture Model and the computation of the probability maps. Finally we show and discuss the results obtained by the different experiments on both datasets.

4.1 Metrics

A way to evaluate the results of a segmentation task is in terms of *precision* and *recall* measures. In this work we use these measures computed at pixel level to evaluate the quality of our methods, being also capable to directly compare with other works in the area.

In a segmentation task, precision is defined as the fraction of detected pixels that are relevant, and recall represent the fraction of correctly detected pixels returned. We also combine both measures to provide the *F1 score*, or *F-measure*, defined as the armonic mean between precision and recall. This score is used in the ICDAR 2009 Page Segmentation Competition together with another metrics to evaluate the presented methods, so we able to directly compare our results on the PRImA dataset. *Precision, recall* and *F-measure* are computed at pixel level as:

$$\text{Precision rate} = \frac{\text{Number of pixels correctly identified}}{\text{Number of pixels detected}}$$

$$\text{Recall rate} = \frac{\text{Number of pixels correctly identified}}{\text{Number of pixels in ground truth}} \tag{4.1}$$

$$\text{F-Measure} = 2 \cdot \frac{\text{Precision } \cdot \text{Recall}}{\text{Precision + Recall}}$$

4.2 Statistical hypothesis test

We are also interested in check whether the inclussion of the Relative Location Features into the model produce significant improvements from a statistical point of view.

A way to perform this verification is computing a Student's t-test on our obtained results, however, as this method is devised to be applied on normal distributions and the F-measure distribution does not satisfy this condition we will compute the Wilcoxon rank-sum test instead. This test is considered the non-parametric version of the Student's t-test, and is suitable for non-gaussian distributions. The test is devised to check whether one of two independent distributions tends to produce larger values than the other. If we apply this verification to the results obtained by our methods, we will be able to check whether the results produced by the RLF approach are significantlly better or not with regard a significance level α.

The statistic U given by the test is computed by the expression:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \tag{4.2}$$

where n_1, in our case, is the number of results obtained only by the Gabor approach on each page of the test set, and R_1 is the sum of the *F-measure* obtained on each image. The statistic U is then given by the minimum value between U_1 and U_2, considering U_2 the same calculation in the case of the results obtained by the RLF approach.

We consider in our experiments a significance level of 5%, which means that values of U under this value will be considered statistically significant.

4.3 Data Partition

Now we describe how the provided datasets are distributed in our training and test partitions. In the case of the 5CofM collection, we use the ground truth generated as was explained in Section 2.2, at the moment of the evaluation of our method we

had a total of 80 pages properly tagged for our task, so we had to stick to these pages.

For the evaluation on the PRImA collection, we used the same test partition provided for the ICDAR competition [1]. The training set was randomly selected from dataset provided in the PRImA website. The following table shows the total number of pages of each collection as well as the used for both training and test partitions.

Dataset	**Total Pages**	**Training**	**Test**
5CofM	80	60	20
PRImA	103	48	55

Table 4.1: Number of pages on each training and test partition for the 5CofM and PRImA datasets

4.4 Experimental settings

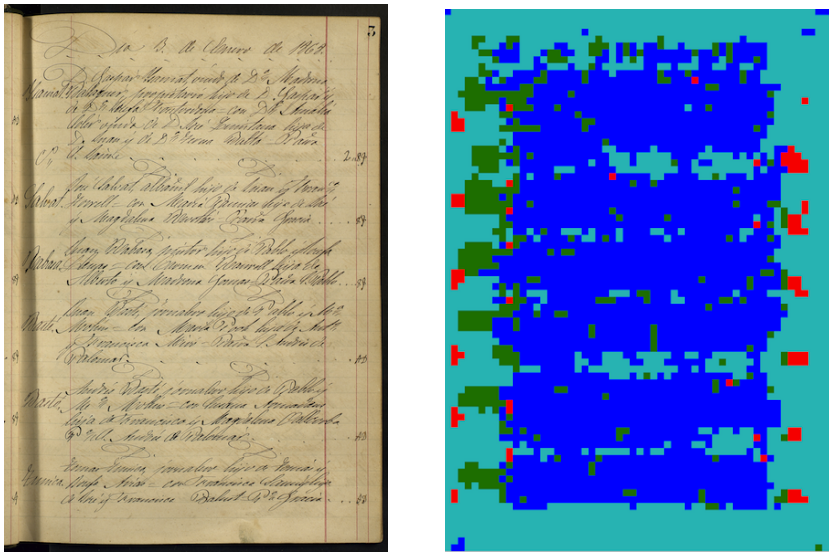
Now we describe the construction process of the Gaussian Mixture Model described in Section 3.4, the parameters and the training process described below was applied equally to both datasets in this work.

The feature extraction process of the images from the training set was performed using the multi-resolution Gabor filter bank proposed by Ilonen *et al.* in [12]. The parameters of the filter bank were manually chosen but ensuring a circular Gabor support and an overlapping degree in the frequency domain of 0.5, we have used a total of 9 orientations in 4 different scales, which leaves a total of 36 different filters in the bank. We also fixed the highest frequency of the filter bank in 0.35 and leave the rest of the filter bank parameters fixed to the default values.

The aplication of the designed filter bank over an image pixel produce a 36-dimensional real-valued feature vector corresponding to the each filter response on that pixel. We apply this process over each image in the training set on a total of 3,000,000 pixels randomly selected. Once we have processed all the images of the training set, we randomly select 500,000 feature vectors corresponding to each class of the dataset. Thus, we use each set of 500,000 samples to learn a gaussian mixture model for each class. We set the number of components of each mixture in 36, the same than the number of components on the feature vector.

Once we have train the model it is possible to obtain the initial cell-level classification. Given an image, we compute our feature vectors on each image pixel, then using one of the equations in section 3.4.2 we assign the corresponding label to each cell.

Figure 4.1 shows the cell-level classification result of one image from 5CofM dataset using Eq. (3.9). We can appreciate as some cells at the left of the image are labeled with the class *tax*, when this class must appears only in the right side



(a) Original image (b) Classification result

Figure 4.1: Initial cell-level classification result of one image from 5CofM dataset

of the image. The same effect can be appreciated for the class *name*, for which we obtain labeled cells wich this class along the entire image. Although the CRF will help to smooth this results in terms of region homogenization, is expected that the inclusion of the Relative Location Features also contribute to correct these mistakes.

In what regards to the implementation of the probability maps, we decided to represent them in a matrix of size 200×200 pixels, so that the computed displacements have to be normalized by the image width and height and quantized to be represented in this range. We also fix the Map coordinates in the range $[-1, 1] \times [-1, 1]$ (see figure 3.3). In addition, each map is normalized in order to define a proper probability distribution and ensure $\sum_{c=1}^{l} M_{c|c'}(u, v) = 1$.

4.5 Cell size experiment

As we introduced in previous chapters, we have divided the image into a set of rectangular cells in order to reduce reduce the number of calculations needed. In this experiment we evaluate the effect of using different cell size in the partition process. We have compared two different cell sizes, 25×25 and 50×50 pixels, as we are interested in reduce to the minimum the number of calculations but ensuring that this process does not affect to the quality of our results.

For the experiment we performed the segmentation of the 20 test images from the 5CofM dataset using the approach composed by both RLF and Gabor features (Eq. (3.14)). The estimation of the initial labelling was performed using Eq. (3.9).

The results of the experiment are shown in Table 4.2. We can appreciate as

Class	Relative Location Features Cell size 25x25			Relative Location Features Cell size 50x50		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Body	0.83	0.96	0.88	0.88	0.96	**0.92**
Name	0.70	0.75	0.72	0.77	0.73	**0.74**
Tax	0.31	0.65	0.41	0.69	0.66	**0.66**

Table 4.2: Results on 5CofM dataset comparing cell sizes of 25×25 and 50×50 pixels.

using cell size of 50×50 we obtain better results in general than using 25×25 pixels per cell. Initially we could have thought that using smaller cell sizes would produce better results, however, the distribution of the classes within the image can explain this results. We can see in Figure 2.2 as the area of the class *tax* is small in comparison with the classes *body and name*, so the division of this region into cells may produce that some *tax* cells are labeled with the class *background*, obtaining a small precision rate as can be seen in the results. The same effect is observed for the class *name* in less proportion.

Considering the obtained results, we will use a cell size of 50×50 pixels for the rest of our experiments.

4.6 5CofM results discussion

Now we show the results obtained in the conducted experiments over the 5CofM dataset. In addition to evaluating the effect of including the Relative Location Features, we also want to check which of the different ways to compute the probability $P(c_k|s_k)$ shown in section 3.4.2 produce better results, as the initial cell classification is a crucial step in the encoding process of the RLF. The results are shown in Tables 4.3, 4.4 and 4.5, which corresponds to Eq. (3.9), Eq. (3.10) and Eq. (3.11), respectively. We show the results achieved for the classes *body, name* and *tax*, as the *background* detection was not of interest in this task. Considering the results, there are some aspects that deserve to be highlighted.

First, we can appreciate that the approach based in counting the number of pixels of each class (Eq. (3.9)) achieves higher rates in terms of F-measure than the other two approaches. Observing the values for the precision and recall we can see as this approach gets lower precision rates, but rather higher in the case of recall, which we consider more valuable for the objectives of this task. Comparing the approaches based in Eq. (3.10) and Eq. (3.11) we can see as the one which takes into account the priori information of the classes gets slightly greater detection rates than the one that not consider this information. This situation can be understood by analyzing the license representation shown in Figure 2.2. We can appreciate that the area belonging to the classes *name* and *tax* represents a small fraction of

the entire license, so the priori probability of those classes is much lower than in classes *body* or *background*. This produce that in the final segmentation a great number of cells are assigned to the wrong class.

	Gabor features			Relative Location Features		
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
Body	0.89	0.87	0.88	0.88	0.96	**0.92**
Name	0.27	0.82	0.40	0.77	0.73	**0.74**
Tax	0.35	0.91	0.50	0.69	0.66	**0.66**

Table 4.3: Precision, Recall and F-measure on 5CofM dataset using Eq.(3.9) for the estimation of $P(c_k|s_k)$.

Secondly, comparing both proposed approaches, we can appreciate in the results as the inclusion of the Relative Location Features permits to obtain higher F-measure rates in the three conducted experiments than using just Gabor features. This increase is mostly produced due to the increment in the precision rate of the classes *name* and *tax* as we predicted before. To check whether the observed result was significant from a statistical point of view we apply the Wilcoxon rank-sum test with a significance level of 5%. The results of the test proved that the improvement seen by the inclusion of the RLF was statistically significant for the three considered classes, so we can affirm that the inclusion of this features helps to improve the segmentation results in this type of documents. This conclusion confirms our initial hypothesis which motivated the realization of this work.

	Gabor features			Relative Location Features		
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
Body	0.89	0.81	0.85	0.88	0.93	**0.91**
Name	0.34	0.23	0.27	0.85	0.35	**0.48**
Tax	0.74	0.27	0.38	0.76	0.44	**0.54**

Table 4.4: Precision, Recall and F-measure on 5CofM dataset using Eq. (3.10) for the estimation of $P(c_k|s_k)$

Finally, we show the result of the segmentation of one page in Figure 4.2. There are some important aspects to highlight from this result. First we observe that, although the body regions are identified with high precision, the different licenses in the page are merged. The space between licenses was identified as a *body* part, which implies that we are not going to be able to directly discriminate between licenses. Secondly, the representation in cells produce that some missing regions specially in the class *tax*, as the major part of the cell is classified as *background* the entire cell is labeled with this class, which influences the recall of this class. Both problems are depicted in Figure 4.3.

Class	Gabor features			Relative Location Features		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Body	0.88	0.74	0.80	0.88	0.94	**0.91**
Name	0.16	0.61	0.26	0.79	0.49	**0.60**
Tax	0.60	0.53	0.54	0.75	0.57	**0.61**

Table 4.5: Precision, Recall and F-measure on 5CofM dataset using Eq. (3.11) for the estimation of $P(c_k|s_k)$

4.7 PRImA Results

Now we show the results of the experiments performed on the PRImA dataset. The methodology of the experiments was the same than in the previous case, we show in Tables 4.7, 4.8 and 4.9 the different results obtained using each approach in 3.4.2. We show in Figure 4.4 the segmentation result on one image from PRImA dataset.

The same result was achieved here respect to the way of computing the probability $P(c_k|s_k)$ than in the previous case. The approach based in counting the number of pixels of each class in the cells gets the higher values of the F-measure and better recall of class *text*. However, the reason of this result is not strictly the same than in the 5CofM dataset. In this dataset there are some images, as images from technical documents, where no elements of the class *image* are found, so the priori probability of each class may be not adequate and representative for all the documents.

In respect to the effect produced by the inclusion of the Relative Location Features into the model, we can appreciate how in this case the improvement achieved is not as large than before. Although we can see a slight increment in the precision detecting both classes and in the final F-measure value, the results obtained by the significance test shows values over the 5%, so the improvement achieved by the RLF can not be considered stadistically significant. Nevertheless, it should be noted that in the case of the class *text* the result of the test is close to be stadistically significant with a *p-value* of 0.05.

Making a comparison of our results with the ones obtained in the ICDAR'09 contest [5] (Table 4.6), we can appreciate as the detection rate in terms of F-measure is up to the highest results achieved in the contest (84%-95%). In the case of our class *image* they do not concretely refer to this class but they consider the class *Non-text* where is included also graphs and other elements, so our results are not directly comparable with them.

Method	Non-text	Text	Overall
DICE	0.66	0.92	0.90
Fraunhofer	0.75	0.95	0.93
REGIM-ENIS	0.67	0.92	0.88
Tesseract	0.74	0.93	0.91
FineReader	0.72	0.93	0.92
OCRopus	0.52	0.84	0.78

Table 4.6: F-measure of the reported methods in ICDAR2009 [5]

	Gabor features			Relative Location Features		
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
Text	0.87	0.91	0.88	0.88	0.94	**0.90**
Image	0.38	0.48	0.42	0.43	0.47	**0.44**

Table 4.7: Precision, Recall and F-measure on PRImA dataset using Eq. (3.9) for the estimation of $P(c_k|s_k)$

	Gabor features			Relative Location Features		
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
Text	0.93	0.56	0.67	0.86	0.73	**0.76**
Image	0.42	0.32	0.35	0.43	0.37	**0.38**

Table 4.8: Precision, Recall and F-measure on PRImA dataset using Eq. (3.10) for the estimation of $P(c_k|s_k)$

	Gabor features			Relative Location Features		
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
Text	0.92	0.60	0.70	0.86	0.74	0.77
Image	0.40	0.29	0.33	0.41	0.35	0.38

Table 4.9: Precision, Recall and F-measure on PRImA dataset using Eq. (3.11) for the estimation of $P(c_k|s_k)$

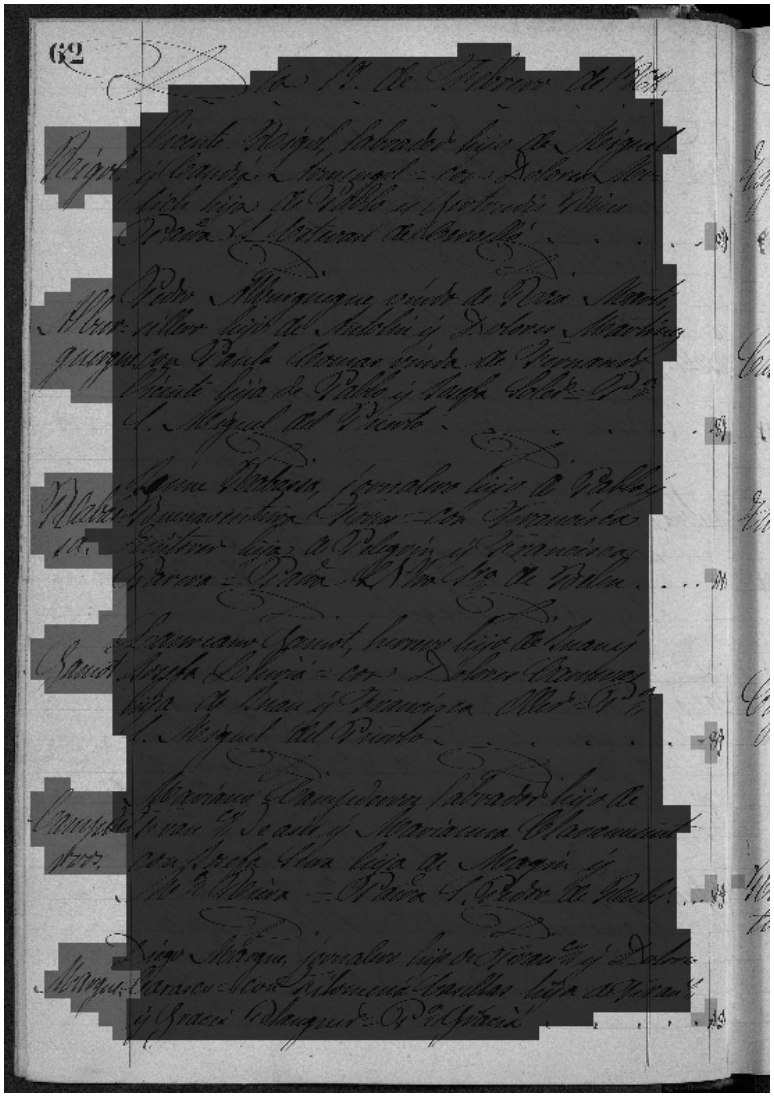


Figure 4.2: Segmentation result of one page from the volume 208 of 5CofM dataset

Figure 4.3: Segmentation results of two licenses of 5Cofm. The red square show an example of missing *tax* regions, and red line indicates the area where the two licenses are merged

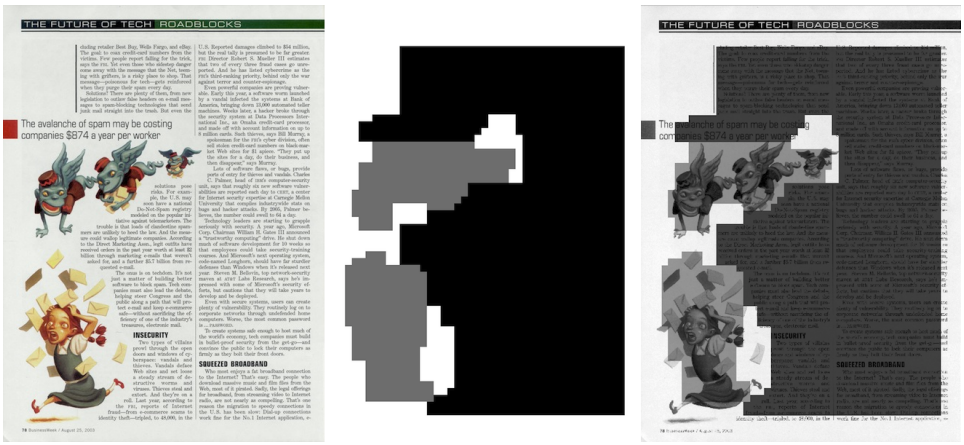


Figure 4.4: Segmentation result of one image from PRImA dataset: Original image (left), segmentation mask (center), overlapped images (right)

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Within the field of Document Image Analysis and Recognition, layout analysis is responsible for analyzing the structure of the document by identifying the location of the different entities that comprise it. According to the distribution of these entities it is possible to discriminate between documents with regular layouts, Manhattan layouts, and documents with an arbitrary distribution along the image, non-Manhattan layouts. In this work we have presented a generic method to perform layout analysis on both types of documents.

Our proposal is focused in texture analysis and the use of Relative Location Features to encode the existent relationships between the different classes of entities. We have built a Conditional Random Field framework that allow us to obtain the best class configuration of an image in terms of energy minimization. In order to evaluate the integration of the Relative Location Features into the model we have proposed two different approaches for the construction of the CRF. In the first place, we use only texture features computed by a multi-resolution Gabor filter bank, then, in second place we add the combine the previous model with the Relative Location Features.

We conducted experiments on two different datasets in order to check the effectiveness in documents with Manhattan and non-Manhattan layouts. The results obtained prove that the inclusion of RLF into the model significantly increase the quality of the segmentation in documents with Manhattan layout and highly structured. In the case of non-Manhatan layout we obtain a slight improvement in the case of detect text regions using RLF, however this improvement is not significantly enough to confirm that the use of RLF helps on this kind of documents. Despite of this, the obtained results on this dataset are up to other state of the art methods in document segmentation, which proves that our method can be applied on both types of layouts with positive results.

5.2 Future Work

During the realization of this work have appeared several situations that have made us think about some improvements that could be used to improve the results, both as regards to the implementation details as some new ideas to improve the quality of the final segmentation of the licenses on 5CofM dataset. In addition, there are several open work lines in the 5CofM project on which we have to continue working.

Regarding to details in the method implementation, we think that improving the inicial cell-level clasification may result in a most appropriate definition of the Relative Location Features. Besides, until now we compute the RLF directly from the classification result just one time, we think that reiterating this process over the values of the RLF in the previous interation it is also possible to improve the RLF accuracy. Another idea in this line is to compute the RLF using the segmentation result obtained by the approach based only in Gabor features instead than on the initial GMM cell-level classification.

As an important part of the 5CofM project, we want to segment the different licenses on each page. As a result of this work the licenses are merging by the class *body* in the region between licenses, we think that this region can be represented separately as another extra class, as it contains some elements that may help to discriminate between this new class and the rest. So, next step will be the integration of this new class in the segmentation process and we expect to obtain in this way a proper license segmentation. In addition, in the future we want to extend our method to the rest of volumes of the 5CofM dataset.

Finally, we are interested in compare the results of our method with other ongoing works in the 5CofM project, concretely in the segmentation of the volume 208, conducted by the PRHLT group from the Univesidad Politécnica de Valencia. For this purpose we have in mind some new evaluation metrics for the result of segmenting the licenses, as computing the relative displacement between the detected bounding boxes with respect to the ground truth.

Appendix A

Contributions

As result of the research process described in this thesis, it has been produced the following publication:

- F. Cruz and O. Ramos Terrades. Document segmentation using Relative Location Features. *21th International Conference on Pattern Recognition 2012 (ICPR2012)*.

Bibliography

[1] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher. A realistic dataset for performance evaluation of document layout analysis. *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009)*, pages 296–300, 2009.

[2] A. Antonacopoulos, B. Gatos, and D. Bridson. Icdar2005 page segmentation competition. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on Document Analysis and Recognition*, pages 75 – 79 Vol. 1, aug.-1 sept. 2005.

[3] A. Antonacopoulos, B. Gatos, and D. Bridson. Icdar2007 page segmentation competition. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 1279 –1283, sept. 2007.

[4] A. Antonacopoulos, B. Gatos, and D. Karatzas. Icdar 2003 page segmentation competition. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, ICDAR '03, pages 688–, Washington, DC, USA, 2003. IEEE Computer Society.

[5] Apostolos Antonacopoulos, Stefan Pletschacher, David Bridson, and Christos Papadopoulos. Icdar 2009 page segmentation competition. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 1370–1374, Washington, DC, USA, 2009. IEEE Computer Society.

[6] Vicente Bosch, Alejandro H. Toselli, and Enrique Vidal. Natural language inspired approach for handwritten text line detection in legacy documents. pages 107–111, 2012.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[8] J. G. DAUGMAN. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, 7:1160–1169, 1985.

[9] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Computer Vision, 2009 IEEE 12th International Conference on Computer Vision*, pages 670 –677, 29 2009-oct. 2 2009.

[10] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *Int. J. Comput. Vision*, 80(3):300–316, December 2008.

[11] Jaekyu Ha, R. M. Haralick, and I. T. Phillips. Recursive x-y cut using bounding boxes of connected components. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 2) - Volume 2*, ICDAR '95, pages 952–, Washington, DC, USA, 1995. IEEE Computer Society.

[12] J. Ilonen, J.-K. Kamarainen, P. Paalanen, M. Hamouz, J. Kittler, and H. Kalviainen. Image feature localization by multiple hypothesis testing of gabor features. *Image Processing, IEEE Transactions on Image Processing*, 17(3):311 –325, march 2008.

[13] J. Ilonen, J.-K. KŁlmŁrŁinen, and KŁlviŁinen H. *Efficient Computation of Gabor Features*. 2005.

[14] A. K. Jain and S. Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Machine Vision Applications*, 5(3):169–184, July 1992.

[15] A. K. Jain, Ratha N. K., and Lakshmanan S. Object detection using gabor filters. *Pattern Recognition*, 30(2):295 – 309, 1997.

[16] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(7):737–747, July 1993.

[17] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi. Text extraction and document image segmentation using matched wavelets and mrf model. *Image Processing, IEEE Transactions on Image Processing*, 16(8):2117–2128, aug. 2007.

[18] Ville Kyrki, Joni-Kristian Kamarainen, and Heikki Kälviäinen. Simple gabor feature space for invariant object recognition. *Pattern Recogn. Lett.*, 25(3):311–318, February 2004.

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[20] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte. Document image segmentation using a 2d conditional random field model. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on Document Analysis and Recognition*, volume 1, pages 407 –411, sept. 2007.

[21] T. Randen and J. H. Husy. Filtering for texture classification: A comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(4):291–310, April 1999.

[22] Y.L. Tong. *The multivariate normal distribution*. Springer series in statistics. Springer-Verlag, 1990.

[23] L. Xingyuan, O. Weon-Geun, J. Soo-Young, M. Kyong-Ae, and K. Hyeon-Jin. An efficient method for page segmentation. In *Information, Communications and Signal Processing, 1997. ICICS*, pages 957 –961 vol.2, sep 1997.