Structure Detection and Segmentation of Documents Using 2D Stochastic Context-Free Grammars

Francisco Álvaro[a], Francisco Cruz[b], Joan-Andreu Sánchez[a], Oriol Ramos Terrades[b], José-Miguel Benedí[a]

[a] *Pattern Recognition and Human Language Technologies, Universitat Politècnica de València*
[b] *Centre de Visió per Computador, Universitat Autònoma de Barcelona*

Abstract

In this paper we define a bidimensional extension of Stochastic Context-Free Grammars for structure detection and segmentation of images of documents. Two sets of text classification features are used to perform an initial classification of each zone of the page. Then, the document segmentation is obtained as the most likely hypothesis according to a stochastic grammar. We used a dataset of historical marriage license books to validate this approach. We also tested several inference algorithms for Probabilistic Graphical Models and the results showed that the proposed grammatical model outperformed the other methods. Furthermore, grammars also provide the document structure along with its segmentation.

Keywords: document image analysis, stochastic context-free grammars, text classification features

1. Introduction

Page segmentation is a fundamental problem of Document Image Analysis (DIA) which is important for solving subsequent document analysis and recognition problems. Document image segmentation intends to detect homogeneous relevant zones in a given document and finding out the structural

Email addresses: falvaro@prhlt.upv.es (Francisco Álvaro), fcruz@cvc.uab.cat (Francisco Cruz), jandreu@prhlt.upv.es (Joan-Andreu Sánchez), oriolrt@cvc.uab.cat (Oriol Ramos Terrades), jmbenedi@prhlt.upv.es (José-Miguel Benedí)

relation among these zones [1]. The relevant zones in DIA depend on the task and they can be drawings, textual zones, special symbols, etc. This paper is focused on determining the structure and the segmentation of textual zones in images of handwritten historical documents. This step is crucial for subsequent text recognition processes.

Many successful image segmentation techniques have been defined in the past for typeset documents [1]. Succesful contests have been held for this type of documents where a common framework is defined in order to be able to compare existing techniques [2, 3]. Many proposed techniques are based on a first step of classification at pixel level, and then a post-processing step where pixels are grouped into regions to obtain uniform zones [4].

In case of historical handwritten documents, the challenge in image segmentation is to detect homogeneous handwritten zones [5, 6]. Correct detection of textual zones is important for tackling subsequent problems like line detection and extraction [7] and later transcription or word spotting [8]. This paper is centered on image segmentation of historical handwritten documents. Developing generic image segmentation techniques for these documents is a very difficult task due to the absence of general editing rules in the past, since the editing rules were usually different for each collection.

Many historical handwritten documents exhibit regularities similar to typeset documents, and image segmentation techniques used for typeset documents can be considered for historical handwritten documents [9]. Segmentation of this kind of documents has been approached in the past with geometrical techniques. In [5] projection profiles were mainly used for page layout analysis of documents with very satisfactory results. But for many other documents, page segmentation techniques that rely on explicit isolation of elements like characters, words or lines are often not useful. For those documents, holistic approaches seem more appropriate. This paper is focused on this second type of historical handwritten documents, concretely in marriage license books [10] (see Figure 1).

Marriage license books are documents that were used for centuries to register marriages in ecclesiastical institutions. Each marriage is represented by a record and the transcription of these documents has been considered very interesting for demography and migratory research [11]. Each unit of information is composed of several related textual regions. Two relevant page segmentation problems can be stated for these documents. First, to segment and classify the different textual units of the records. And second, to find out the syntactic structure of the records in a given page.

Probabilistic graphical models (PGM) offer a natural framework to tackle these segmentation problems and to relate segmented units represented here as random variables, since it easily allows to represent dependencies between them [12, 13, 14]. However, computing exact inference on these models may be challenging depending on the structure that they present. In this case we must resort to other approximate methods like the Graph Cut algorithm [15] or some variations of the Belief Propagation (BP) algorithm [16]. Within this formal framework, in [17] a solution is proposed for classifying the different textual zones that are present in marriage license books, although no structure detection is performed. In that research, pixel classification based on texture features obtained from the Gabor transform are compared with Relative Location Features [18]. Both sort of features are combined in a Conditional Random Field [19] to take into account contextual information in the classification process of the pixels.

In order to address both the detection of textual zones and the analysis of structural relationships among these zones, we consider the use of structural models, such as Stochastic Context-Free Grammars (SCFG). SCFG are a powerful formalism of Syntactic Pattern Recognition which has been used previously for Document Image Analysis [20, 21]. Bidimensional SCFG (2D-SCFG) is a well known formalism that has been studied in the past for bidimensional parsing [22, 23]. This type of grammars are able to represent efficiently contextual bidimensional relations that are important for page segmentation [24]. In this study we propose a formal model that integrates several stochastic models for textual zone segmentation and structural analysis directly into the parsing process of 2D-SCFG. The contributions of this paper with respect to [24] are the following. This paper researches the probabilistic estimation of the grammatical models. We compare additional approaches and we use a larger dataset that allowed us to carry out a more comprehensive experimental research. Moreover, we provide a more detailed description of the methodology used with 2D-SCFG.

In the following Section 2 we describe the problem of structure detection and page segmentation applied to marriage license books. A review of PGMs is given in Section 3. The 2D-SCFG model and the corresponding parsing algorithm are defined in Section 4. Then we describe the features used for classifying textual zones at local level in Section 5. Finally, Section 6 reports and analyzes the experimentation carried out, and conclusions and future work are provided in Section 7.

2. Segmentation of Structured Documents

Marriage license books are handwritten documents that have been used in ecclesiastical institutions for centuries for registering marriages. Most of these books have a structure similar to an accounting book. Figure 1 shows an example of page of a marriage license book belonging to a collection of 291 books conserved at the Cathedral of Barcelona. The pages in these books were orderly written, and although there are differences over the centuries, the layout in each page was quite rigid.

Every book is divided in two parts: the first part is an index of surnames and the second part contains the marriage license records (see [10] for a more detailed description of this collection). This paper is focused on the segmentation of the pages in the second part of the book.

Each page contains several records, such that each one is associated with a marriage license. Each record has in turn a *husband surname's block* (Figure 2.a), the *main block* (Figure 2.b), and the *tax block* (Figure 2.c). Note that the documents can have additional textual zones, like the date that can be seen at the beginning of the page (it can also appear in the middle of a page), and the two large calligraphic letters[1] that separate the consecutive records that were registered the same day. These additional zones were ignored in this paper, i.e., they are considered like background because they were not considered relevant for subsequent transcription tasks. The process for creating the ground-truth requires marking the minimum rectangle containing the identified classes: *Body*, *Name* and *Tax*. All the pixels that did not belong to any of these regions were considered background.

The final goal in these documents is to obtain the transcription of each marriage license. The transcription of a similar document was studied in [10] by using Handwritten Text Recognition (HTR) techniques [8]. In that paper, HTR experiments were carried out by using lines as the minimal unit segmentation for training and recognition. Using lines for this purpose has the drawback that there is no context for the language model at the beginning of the line and most of the errors are usually concentrated in the initial words of the line. Therefore, concatenating the lines of a record should be very important to transcribe this type of documents at record level, as well as other techniques like category-based language models [25].

[1] These letters are D. D. that is the abbreviation of "Dit dia" which means "The mentioned day".

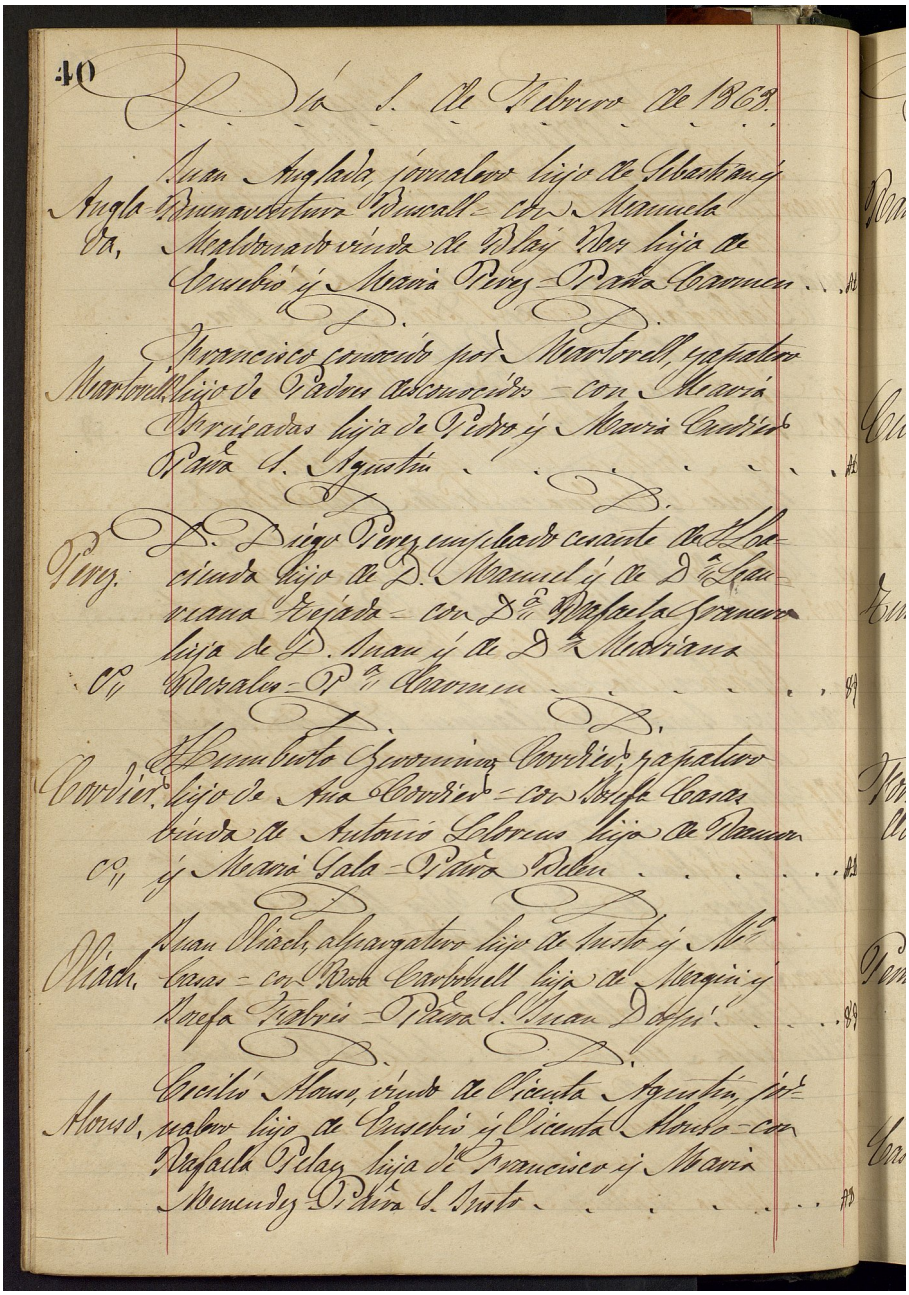


Figure 1: Example of page of a marriage license book containing six records.

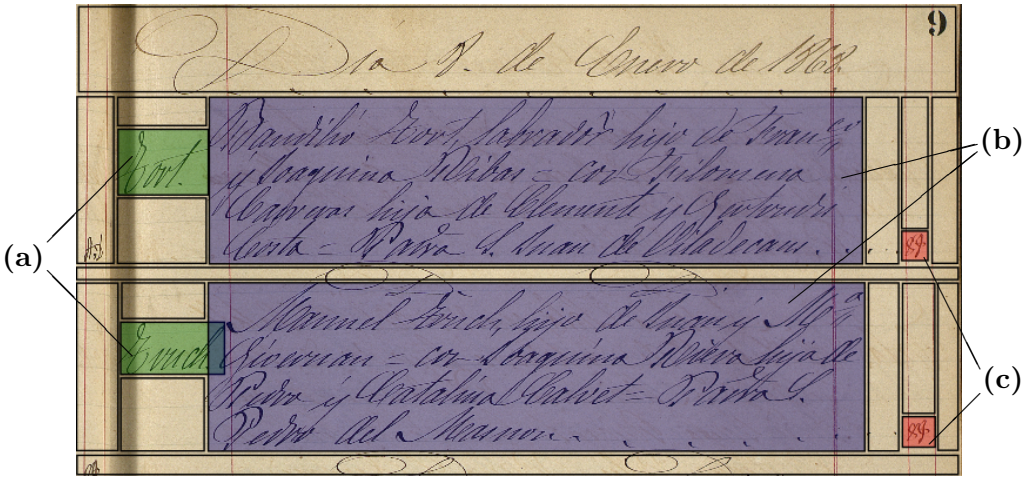


Figure 2: Example of the page segmentation problem for two records. Several background zones are considered and each record is composed of three parts: (a) Name (b) Body (c) Tax.

In this way the correct segmentation of each page becomes a challenging task. The problem is to correctly isolate every record in a page, and to relate their corresponding parts, that is, the surname, the body text and the tax associated with each entry. In this paper we focused on detecting the bounding boxes around the main parts of each record. Note, that a fine-grained detection of the frontiers of each zone would be ideal, but this is difficult because sometimes two zones overlap if rectangular bounding boxes are used as in this paper (see the lower record in Fig. 2).

The problem of detecting the records can be stated as two different problems: first, to classify the textual zones into the previously mentioned classes (*Background*, *Name*, *Body* and *Tax*); and second, to detect the complete set of records of each page. To address this problem, first we review related work based on PGMs to solve the segmentation of images of documents. These graphical models become our baseline approach. Second, we present a model based on 2D-SCFG that solves the segmentation of the full document using structural and stochastic information. Finally, we describe two sets of text classification features used to classify the image regions according to the graphical information, and the experimentation performed on this corpus.

3. Probabilistic Graphical Models

When we tackle the problem of image labeling, Probabilistic Graphical Models (PGMs) [14] provide a proper framework to represent the structure and the relationships between the variables in the model.

In this representation the different variables are distributed in a graph structure, where it can be depicted as a directed or undirected graph depending on the type of dependencies represented. This graph is composed by a set of nodes representing the different set of variables, and a set of edges denoting the dependencies between the nodes.

In the case of image labeling, a natural way for representing the dependencies between the pixels of the image is by means of a bidimensional grid-like structure, which can be modeled by a Markov Random Field (MRF) [26]. In this representation each pixel in the image is represented by a node in the graph, although in some tasks it is also common that a node represents a group of pixels clustered in cells or superpixels [18].

In this problem the objective is to compute *Maximum a Posteriori* (MAP) probability to find the combination of class labels c for each pixel in x that maximizes the PGM probability. One way to model this distribution is in terms of the energy associated with a Conditional Random Field (CRF) [19], conditioning the probability with respect to a set of computed features:

$$P(c \mid x) = \frac{1}{Z(x)} \exp\left\{ \sum_i \psi(x_i, c_i) + \sum_{(i,j)\in\varepsilon} \phi(c_i, c_j) \right\} \tag{1}$$

where $\psi(x_i, c_i)$ represents the local potentials at each pixel i, and $\phi(c_i, c_j)$ the pairwise potentials of assigning the labels c_i, c_j to the neighbor pixels i, j. The constant $Z(x)$ represents the partition function, a normalization factor to ensure the proper definition of the probability distribution. In some types of structures, as in the case of grid-like graphs, a large amount of variables may result in the impossibility of providing an exact computation of this value, leading to the need of using approximate methods to achieve this task [27].

Many methods have been proposed to perform inference in PGMs, that is, to obtain the likelihood or the conditional probability with a model for a given input. However, the problem of computing exact inference in grid-structured CRFs is known to be a NP-hard problem and becomes intractable when we

model a large number of variables [28]. Nevertheless, there are many methods in the literature that provide approximate solutions to the problem. One example is the Graph Cut algorithm [15] used in many segmentation tasks [29]. The method relies on the fact that many computer vision problems can be formulated in terms of an energy minimization function, and provides a local minimum of the minimization function based on the most likely cut in the graph. Another family of methods used to perform approximate inference in graphical models are the sum-product message passing algorithms. Within this type we found the Belief Propagation (BP) algorithm [16], and the version for loops Loopy Belief Propagation (LBP) [30], that is able to perform inference in the case of grid-structured CRFs. There are also other algorithms that follow different approaches. One example is the Iterated Conditional Models (ICM) [31], an algorithm for optimization that follows a search paradigm. In this paper we use the three algorithms stated before for the inference in PGMs.

4. 2D Stochastic Context-Free Grammars

In this study we propose to use 2D-SCFG in order to compute the most likely structure and segmentation of a document. This powerful model intends to tackle the logical layout problem in combination with text classification features. A context-free model is a natural way to account for both the horizontal and vertical context of the problem, where there are dependencies among rows, columns and 2D regions.

We formally define a 2D-SCFG as follows. A *Context-Free Grammar* (CFG) G is a tuple (N, Σ, S, R), where N is a finite set of non-terminal symbols, Σ is a finite set of terminal symbols ($N \cap \Sigma = \emptyset$), $S \in N$ is the start symbol of the grammar, and R is a finite set of rules: $A \to \alpha$, $A \in N$, $\alpha \in (N \cup \Sigma)^+$.

A *Stochastic Context-Free Grammar* (SCFG) $\mathcal{G}_s$ is defined as a pair (G, P), where G is a CFG and $P : R \to]0, 1]$ is a probability function of rule application, i.e. $\forall A \in N : \sum_{i=1}^{n_A} P(A \to \alpha_i) = 1$; where n_A is the number of rules associated with non-terminal symbol A. This type of grammars can be represented in Chomsky Normal Form (CNF) resulting in only two types of productions: binary rules $A \to BC$ and terminal rules $A \to c$ (where $A, B, C \in N$ and $c \in \Sigma$).

We define 2D-SCFG that are able to deal with bidimensional matrices as a generalization of SCFG. In this extension, nonterminal symbols account for

2D regions. The binary rules of a 2D-SCFG have an additional parameter $r \in \{\mathrm{H}, \mathrm{V}\}$ that describes a spatial relation: horizontal concatenation (H) or vertical concatenation (V). Given a rule $A \xrightarrow{r} B\ C$, the combined subproblems B and C must be arranged according to the spatial relation constraint, i.e., horizontally adjacent and same height for $r = \mathrm{H}$ and vertically adjacent and same width for $r = \mathrm{V}$. This simple extension is enough to account for the problem we are dealing with. The segmentation of the input document can be obtained as the most likely derivation given a 2D-SCFG, such that the region that defines the input image is recursively divided either vertically or horizontally into smaller rectangular regions.

4.1. Parsing Algorithm

Given a page image, the problem is to obtain the most likely parsing according to a 2D-SCFG. For this purpose, the input page is considered as a bidimensional matrix I with dimensions $w \times h$ and each cell of the matrix can be either a pixel or a cell of $d \times d$ pixels. Then, we define an extension of the well-known CYK algorithm to account for bidimensional structures. We have basically extended the algorithm described in [22] to include the stochastic information of our model.

The CYK algorithm for 2D-SCFG is essentially a *dynamic programming* method, which fills in a parsing table $\mathcal{T}$. Following a notation very similar to [32], each element of $\mathcal{T}$ is a probabilistic nonterminal vector, where their components are defined as:

$$\mathcal{T}_{(x,y),(x+1,y+1)}[A] = \hat{P}(A \Rightarrow z_{(x,y),(x+1,y+1)}) \quad (2)$$

$$\mathcal{T}_{(x,y),(x+i,y+j)}[A] = \hat{P}(A \overset{+}{\Rightarrow} z_{(x,y),(x+i,y+j)}) \quad (3)$$

Each region $z_{(x,y),(x+i,y+j)}$ is defined as a rectangle delimited by its top-left corner (x, y) and its bottom-right corner $(x + i, y + j)$. We denote $\ell_{i\times j} = \ell(z_{(x,y),(x+i,y+j)})$ as the size $(i \times j)$ of the subproblem associated with a region $z_{(x,y),(x+i,y+j)}$. The probabilities $\hat{P}$ represent the probability of the most likely derivation from nonterminal A resulting in the region z.

If the size of the subproblem is larger than 1×1, then there exists some binary rule ($A \xrightarrow{r} B\ C$, with $B, C \in N$, and $r \in \{H, V\}$) and some split point k such that, in a similar way to [32], we can divide the problem in two

subproblems:

$$\begin{aligned}\hat{P}(A \overset{+}{\Rightarrow} z_{(x,y),(x+i,y+j)}) \;&=\; P(\ell_{i\times j} \mid A) \;\max_{B,C}\; \{ \\ &\max_{1\le k<i} P(A \xrightarrow{H} B\,C)\; \hat{P}(B \overset{+}{\Rightarrow} z_{(x,y),(x+k,y+j)})\; \hat{P}(C \overset{+}{\Rightarrow} z_{(x+k,y),(x+i,y+j)})\;, \\ &\max_{1\le k<j} P(A \xrightarrow{V} B\,C)\; \hat{P}(B \overset{+}{\Rightarrow} z_{(x,y),(x+i,y+k)})\; \hat{P}(C \overset{+}{\Rightarrow} z_{(x,y+k),(x+i,y+j)})\;\} \quad (4)\end{aligned}$$

where a new hypothesis is computed from two smaller subproblems, such that the probability is maximized for every possible vertical and horizontal decomposition resulting in the region $z_{(x,y),(x+i,y+j)}$. It should be noted that the 2D-SCFG provides syntactic and spatial constraints $P(A \xrightarrow{\mathrm{r}} B\,C)$, and we have also included the probability $P(\ell_{i\times j} \mid A)$ that a nonterminal A accounts for a problem of size $i \times j$.

The probability $P(\ell_{i\times j} \mid A)$ has two effects on the parsing process. First, it helps to model the spatial relations among every part of a given page because there is a specific nonterminal for each zone of interest. For instance, this can be seen in Figure 2 where the size of the background region on top of the page will be different from the size of the background zone over a tax region. Furthermore, many unlikely hypotheses are pruned during the parsing process due to its size information, hence, it speeds up the algorithm.

Considering the definition of the matrix parsing table $\mathcal{T}$ (Eq. (3)), the expression of the Eq. (4) can be rewritten to obtain the general term of the parsing algorithm. Thus, for all i and j, $2 \le i \le w,\ 2 \le j \le h$, we have:

$$\begin{aligned}\mathcal{T}_{(x,y),(x+i,y+j)}[A] \;&=\; P(\ell_{i\times j} \mid A) \;\max_{B,C}\; \{ \\ &\max_{1\le k<i}\; P(A \xrightarrow{\mathrm{H}} B\,C)\; \mathcal{T}_{(x,y),(x+k,y+j)}[B]\; \mathcal{T}_{(x+k,y),(x+i,y+j)}[C]\;\;, \\ &\max_{1\le k<j}\; P(A \xrightarrow{\mathrm{V}} B\,C)\; \mathcal{T}_{(x,y)(x+i,y+k)}[B]\; \mathcal{T}_{(x,y+k),(x+i,y+j)}[C]\;\;\}\end{aligned}$$

For subproblems of size equal to 1×1 and taking into account the definition of Eq. (2), the derivation probability of a single cell (size region equal to 1×1) can be marginalized according to the class label (terminal) c. Given that we need to calculate the most likely parsing, we can approximate the sum by a maximization, and considering some other usual assumptions the probability of the derivation of a single cell is:

$$\hat{P}(A \Rightarrow z_{(x,y),(x+1,y+1)}) \approx P(\ell_{1\times 1} \mid A) \;\max_{c}\; P(A \to c)\; P(c \mid z) \qquad (5)$$

where $P(\ell_{1\times 1} \mid A)$ is the probability that nonterminal A derives a subproblem of size 1×1; $P(c \mid z)$ represents the probability that a cell (region) z belongs to class c, and it is described in Section 5; and $P(A \to c)$ is the probability of a terminal rule for terminal (class) c.

Taking into account the matrix $\mathcal{T}$ (Eq. (2)), we can rewrite the expression of Eq. (5) to obtain the initialization term of the parsing algorithm. Thus, for each region z of size 1×1, we have:

$$\mathcal{T}_{(x,y),(x+1,y+1)}[A] \;=\; P(\ell_{1\times 1} \mid A) \max_{c} P(A \to c)\, P(c \mid z_{(x,y),(x+1,y+1)}) \quad (6)$$

Finally, the most likely parsing of the full input page is obtained in $\mathcal{T}_{(0,0),(w,h)}[S]$ such that S is the start symbol of the 2D-SCFG. It is important to notice that all the probability distributions involved in the parsing process can be learnt from labeled data. The time complexity of the algorithm is $O(w^3h^3|R|)$ and the spatial complexity is $O(w^2h^2)$.

4.2. Model Estimation

The model based on 2D-SCFG for parsing structured documents has, in turn, some stochastic distributions that need to be learnt. First, the probability $P(c \mid z)$ that a certain region z of the image belongs to class c is described in Section 5.

There are two additional distributions that we have to estimate: the probabilityies $P(\ell_{i\times j} \mid A)$ and the probability of the rules of the grammar $P(A \to \alpha)$. In order to learn automatically these distributions, we performed a forced recognition of the training set. Given a certain document, the forced recognition was carried out by providing the probability $P(c \mid z)$ using the ground-truth information. Concretely, for each cell z belonging to class c^* we set $P(c^* \mid z) = 1$ and $P(c \mid z) = 0,\ \forall c \neq c^*$. The remaining distributions were considered equiprobable.

Hence, we obtained for each document the best parsing according to the 2D-SCFG model. On one hand, the probability distribution of the size for each nonterminal A was estimated according to the occurrences in the forced recognition of the training set as

$$P(\ell_{i\times j} \mid A) = \frac{n(A_{i\times j})}{n(A)}$$

such that $n(A_{i\times j})$ is the number of times that nonterminal A accounts for a region of size $i \times j$ in the training set, and $n(A)$ the total number times that nonterminal A accounted for a region of any size.

On the other hand, the probabilities of the rules of the grammar were estimated using the set of derivation trees obtained from the forced recognition of the training set as:

$$P(A \to c) = \frac{n(A \to c)}{n(A)}$$

$$P(A \xrightarrow{r} B\,C) = \frac{n(A \xrightarrow{r} B\,C)}{n(A)}$$

where each rule probability is computed using the number of times that the rule was used in the training set, normalized by the total number of rules with nonterminal A as left-hand symbol $n(A)$.

In order to make the model able to account for unseen events, after these distributions were estimated, we also smoothed them by setting a minimum probability threshold.

5. Text Classification Features

In this section we describe the different features selected in this paper for classifying small regions of pixels and how we incorporated them into the 2D-SCFG described above. Following the outline in [17] we used two different set of features, Gabor features as texture descriptors and Relative Location Features [18].

5.1. Texture features

Gabor transform is a multi-resolution transform commonly used for texture analysis. A bank of filters is defined for several orientations and signal frequencies. A fast implementation of this filter was proposed in [33] and it implements a multi-resolution bank of filters in which the only parameters to be given are: the number of orientations (n), the number of resolution levels (m) and the highest frequency (f_{max}). As a result, it is obtained a feature vector g of dimensions $n \times m$, which covers almost all the spectrum of frequencies up to the highest one f_{max}.

The Gabor filter is defined by a sinusoidal wave of complex values modulated by an exponential function [34]. This exponential function is a Gaussian function centered in the origin of coordinates, with a parameter controlling the size of the function support[2]. In the frequency space, the Gabor filter is

[2]We refer as support of a Gaussian function the region enclosing 99% of the energy.

also defined by a Gaussian function, centered in the frequency f_0 and support inversely proportional to frequency f_0. Furthermore, in images the filter support has an elliptical shape tuned by three parameters γ, η and θ:

$$\begin{aligned} \psi(x, y; f, \theta) &= \frac{f_0^2}{\pi\gamma\eta} e^{-\left(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\eta^2}y'^2\right)} e^{i2\pi f x'} \\ x' &= x\cos\theta + y\sin\theta, \\ y' &= -x\sin\theta + y\cos\theta \end{aligned} \tag{7}$$

It is well-known that the Fourier transform of a Gaussian function is again a Gaussian function. In addition, if we scale the support of Gabor filters by a factor of k^{-m}, the support of their Fourier transform are proportional to k^m. In particular, given the definition of Gabor filters in Eq. (7), the support of Gabor filters in the spatial domain are ellipses with axis proportional to $\frac{\gamma}{f_{max}}k^m$ and $\frac{\eta}{f_{max}}k^m$. The values of η and γ are obtained according to the number of orientations n, the scaling factor k, and the overlapping degree q of filters in the Fourier space as:

$$\gamma = \frac{k-1}{k+1}\frac{\sqrt{-\log q}}{\pi}; \quad \eta = \frac{\sqrt{-\log q}}{\pi\tan\frac{\pi}{2n}}$$

Once we obtained the set of features, we applied a Gaussian mixture model (GMM) to estimate the probability $P(g \mid c)$ of each possible class c identified in this task: *Background*, *Name*, *Body* and *Tax* . Finally, we defined the probability $P(c \mid z)$ for a cell z for a particular class c required in Eq. (6) in the case of the 2D-SCFG, and to define the term ψ in Eq. (1):

$$P(c \mid z) = \frac{1}{|z|}\sum_{g\in z}\frac{P(g \mid c)\,P(c)}{\sum_c P(g \mid c)\,P(c)} \tag{8}$$

5.2. Relative Location Features

Relative Location Features (RLF) were introduced in [18] as a way to encode inter-class and intra-class spatial relationships as local features. These features are computed from *relative location probability maps* $\mathcal{M}_{c_i|c_j}$, encoding the probability of the class c_i at region i and knowing that at region j the class c_j is found. In other words, $\mathcal{M}_{c_i|c_j}(u_{i,j}) = P(c_i \mid c_j, i, j)$ where by $u_{i,j} = (x_i, y_i) - (x_j, y_j)$ we denote the offset between regions i and j, such

that (x_i, y_i) and (x_j, y_j) represent the centroid coordinates of regions i and j, respectively. Thus, the *self* and *other* RLF are defined as:

$$v_{c_i}^{other}(i) = \sum_{i \neq j : c_i \neq c_j} \mathcal{M}_{c_i|c_j}(u_{i,j}) P(c_j \mid j)$$

$$v_{c_i}^{self}(i) = \sum_{i \neq j : c_i = c_j} \mathcal{M}_{c_i|c_j}(u_{i,j}) P(c_j \mid j)$$

Moreover, herein $c_j = \arg\max_c P(c \mid j)$ is the class label assigned at region j having the highest probability and $P(c \mid i)$ is the *a posteriori* probability estimated of Eq. (8). Each of these features, $v_{c_i}(i) = (v_{c_i}^{other}(i), v_{c_i}^{self}(i))$, model the probability of assigning the class label c to a region z taking into account the information provided by the rest of image regions about their position and its initial label predictions. Finally, once we have computed the set of RLF, we are able to compute the probability required in Eq. (6):

$$P(c \mid z) = w^{app} \log P(c \mid z) + w_c^{other} \log v_c^{other}(z) + w_c^{self} \log v_c^{self}(z) \qquad (9)$$

where v_c^{other} and v_c^{self} are the different sets of RLF and w^{app}, w_c^{other} and w_c^{self} are the corresponding weights learnt from a logistic regression model. Further details about the process can be seen in [17].

6. Experiments

This section describes the experimentation carried out to evaluate the proposed 2D-SCFG model for the task of page segmentation. We compare the results obtained with approaches based on PGMs using three different families of inference methods.

First, we describe the used dataset and the general settings of the experiments performed. Then we describe the general outline from both experiments. Finally, we report several performance metrics and the discussion comparing the different models.

6.1. Experimental Settings

The Five Centuries of Marriages (5CofM) dataset is composed of a set of 291 handwritten books including marriage records conducted in the period from the year 1451 until 1905. The set of books includes approximately $550,000$ marriage licences from 250 parishes [10]. Despite the great number

of volumes in this dataset, currently only a few of them are being used on different tasks like handwriting recognition, word spotting, or layout analysis. For each of these tasks the corresponding ground-truth was manually obtained by selecting and labeling the different regions on each page, which is a time consuming process. Therefore, due to time limitations we focused on a particular book of the collection for the experiments reported in this paper. However, as the documents in all the volumes have the same structure, the proposed model could be applied to the remaining books.

This paper is focused on the segmentation of volume 208 of this collection (Figures 1 and 2 show examples). This volume has 593 pages of which we labeled at pixel level the first 200 pages of the volume. We randomly split 150 pages for training, 10 pages as validation set and the remaining 40 for test. The resolution of the images is 300 dpi ($\approx 2750 \times 3940$ pixels).

Following previous works [17, 24] each page of the dataset was divided in cells of 50×50 pixels in order to reduce the computational cost of processing an image at pixel level. In previous studies we tested several configurations of cell sizes and the impact on the results. The experiments using cells of size 25×25 pixels produced lower precision and recall values for all the considered classes in this task. Smaller cell sizes produce that text classification features do not have enough information in order to correctly discriminate among the different classes. Also, cells over 50×50 pixels were not tested since the regions for some of the classes could be smaller than the cell size and they might not be detected.

With respect to the parameters of the texture filter bank, we computed a 36-dimensional feature vector using 9 orientations and 4 frequencies of the filter. These values were chosen to ensure that the Gabor functions cover the frequency space. Additionally, we set the overlapping degree $q = 0.5$, $f_{max} = 0.35$, and the scaling factor $k = \sqrt{2}$. Finally, we also learnt the parameters of the logistic regression used in Eq. (9) from the training set.

6.2. PGMs Experiments

We performed several experiments using the CRF model defined in Eq. (1) to compute the best label configuration for a page. We used the two sets of features described in the previous section to define the local potentials on each cell of the image as described in Eq. (8) and (9). To compute the values for the pairwise potentials we learnt the frequencies for each possible pair of classes from the training set.

We conducted several experiments using different inference algorithms for the CRF. First, we tested the α-β-*swap* version of the Graph Cut algorithm proposed in [15]. Second, we use the method based in message passing LBP [30]. Finally we used the ICM algorithm based in the search paradigm [31].

6.3. 2D-SCFG Experiments

We used a 2D-SCFG to tackle the page segmentation problem on the 5CofM dataset. Given the nature of the problem such that the documents have a known structure, we manually defined the grammar. According to the model described in Section 4, we had to train several probability distributions. The probabilities of the productions of the grammar and the size probabilities for each nonterminal were estimated from the training data as explained in Section 4.2.

The 2D-SCFG model combines probability distributions that were learnt independently, hence, there may be scaling problems when multiplying the different probabilities. For this reason, the resulting probability was obtained such that each distribution had an exponential weight that adjusted the scale of them. As a result, we had to tune three weights: the probabilities of the grammar $P(A \rightarrow c)$ or $P(A \xrightarrow{r} B\,C)$, the probability of a region $P(c \mid z)$ and the probabilities $P(\ell_{i\times j} \mid A)$. Then, the weights of the system were tuned using the downhill simplex algorithm by maximizing the average F-measure for classes *Name*, *Body* and *Tax* when recognizing the validation set.

6.4. Results and Discussion

We classified the document images of the test set by using the model learnt from the training set and the best parameters of the validation experimentation. We tested in these experiments the performance of three inference algorithms for PGMs (Section 3): Graph Cut, Loopy Belief Propagation (LBP) and Iterated Conditional Models (ICM); and a grammatical model (Section 4): 2D Stochastic Context-Free Grammars (2D-SCFG). These models were evaluated in combination with two sets of features which were described in Section 5: Gabor features and Relative Location Features (RLF). Table 1 shows the results for each class: *Body*, *Name* and *Tax*. The reported metrics are the precision, recall and F-measure at cell level averaged for each page in the test set. Results show that class *Body* was classified with good F-measure rates, whereas classes *Name* and *Tax* represented the most challenging part. This is related with the percentage of the page that

Table 1: Classification results for different models and text classification features.

Model	Features	Class	Precision	Recall	F-measure
Graph Cut	Gabor	Body	0.91	0.79	0.84
		Name	0.21	0.80	0.32
		Tax	0.50	0.84	0.61
	RLF	Body	0.90	0.92	0.91
		Name	0.66	0.78	0.70
		Tax	0.89	0.43	0.56
LBP	Gabor	Body	0.89	0.83	0.86
		Name	0.27	0.74	0.39
		Tax	0.45	0.79	0.56
	RLF	Body	0.88	0.93	0.90
		Name	0.72	0.71	0.69
		Tax	0.85	0.43	0.55
ICM	Gabor	Body	0.89	0.83	0.85
		Name	0.27	0.74	0.39
		Tax	0.45	0.79	0.56
	RLF	Body	0.89	0.92	0.91
		Name	0.71	0.74	0.72
		Tax	0.90	0.45	0.58
2D-SCFG	Gabor	Body	0.91	0.95	0.93
		Name	0.73	0.86	0.78
		Tax	0.69	0.80	0.71
	RLF	Body	0.90	0.95	0.92
		Name	0.77	0.79	0.77
		Tax	0.78	0.65	0.68

represents each region, because it is more difficult to properly classify small regions. Errors are usually made in the boundaries of the regions, hence, a row or column of cells represents a smaller percentage of class *Body* than class *Name*, and of course than class *Tax* which is usually composed of just a few cells.

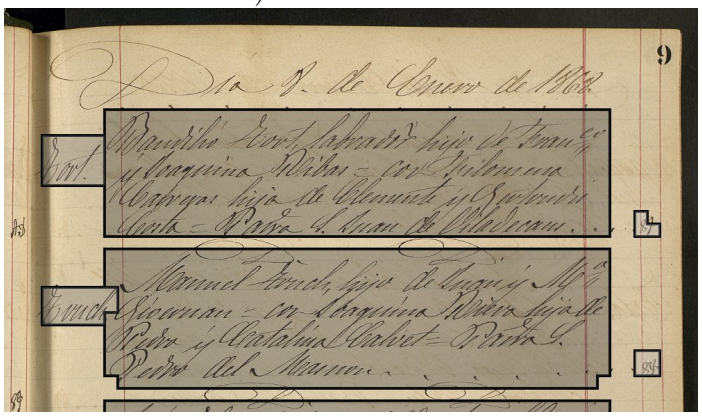
There are two factors to take into account: the text classification features and the page segmentation model. Regarding text classification features, we can clearly observe two different behaviours. On one hand, PGMs performed significantly better with RLF features than when regular Gabor features were used. Both *Body* and *Name* recognition classes always improved, where the improvement in *Name* F-measure is remarkable. The class *Tax* was the most challenging class. Although the differences in F-measure were small, we can see that Gabor features provided less precision but more recall whereas RLF produced higher precision and lower recall values.

On the other hand, the 2D-SCFG model obtained similar performance using both sets of features, and even results with Gabor features were slightly better than results provided by RLF features. 2D-SCFG is a powerful model that is able to take advantage of the knowledge about the document structure. Thus, grammars were able to overcome the lacks of Gabor features obtaining very good results for all classes without the additional spatial information provided by RLF. Given that we learnt stochastic information about the structure of the documents from training data, the model was able to successfully parse using the regular Gabor features. Figure 3 shows an example of recognition using 2D-SCFG and both sets of features. We can see how the overlapping region between *Name* and *Body* is classified as *Body*. Also, as results pointed out (see Table 1), Gabor features obtained higher recall and we can see that resulting regions are larger. On the other hand, RLF provided higher precision by adjusting better the size of the regions detected. Finally, recognizing the space between records is difficult due to the large calligraphic letters considered as background.

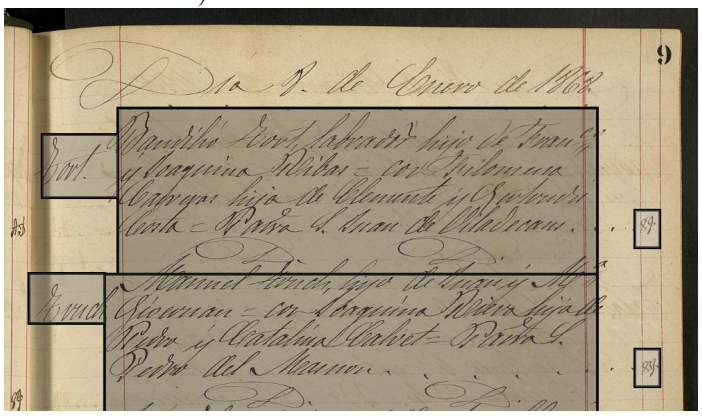
Comparing the performance of the different models, 2D-SCFG significantly outperformed PGMs. Results showed that grammars achieved a great improvement even using Gabor features. The best results among the PGMs were obtained by ICM and RLF features with F-measure 0.91, 0.72 and 0.58 for classes *Body*, *Name* and *Tax*, respectively. The 2D-SCFG model with Gabor features achieved F-measure 0.93, 0.78 and 0.71 for classes *Body*, *Name* and *Tax*, respectively.

PGMs classify cells but they do not identify the explicit segmentation

a) Ground-truth



b) 2D-SCFG with Gabor



c) 2D-SCFG with RLF

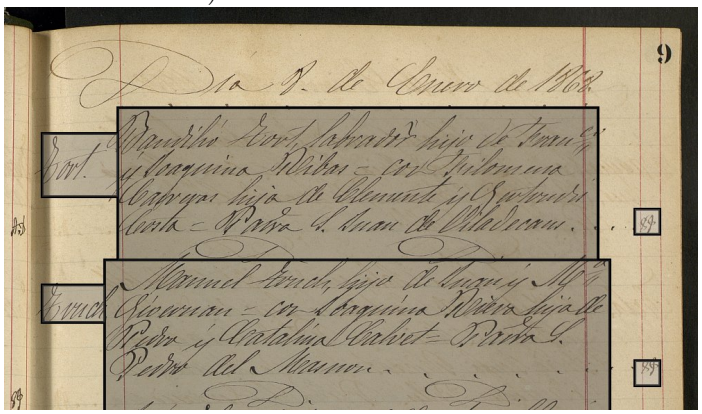


Figure 3: Example of page segmentation and structure detection with 2D-SCFG using cells of 50×50 pixels and different text classification features.

in records. However, the most likely hypothesis according to the 2D-SCFG model provides a derivation tree that accounts for both the structure of the document and the segmentation in cells. Using this information we extracted the number of records detected in each document. Then, we computed the percentage of documents in the test set where the number of records detected was correct.

2D-SCFG in combination with Gabor features computed the right number of records in 80% of the test documents, whereas with RLF features only 52.5% of the documents had the correct number of records detected. All the errors were due to oversegmentation in both sets of features. This measure helps to assess the quality of the recognition such that we can see that in addition to the slight improvement of Gabor features with respect to RLF features, the number of records detected presented an important difference.

7. Conclusions and future work

In this paper we have proposed a model based on 2D-SCFG for page segmentation of structured documents using two sets of features: Gabor and RLF. We also tested several inference algorithms for PGMs, where RLF features obtained better results than Gabor features. The experimentation carried out proved that 2D-SCFG significantly outperformed PGMs in this task. Furthermore, 2D-SCFG obtained better results with Gabor features than using RLF features. Moreover, grammars were able to provide the detailed and explicit information of the page segmentation, hence, record-level evaluation could be done and results also showed a good performance of the model.

Future work will be focused in testing other text classification features that could improve the segmentation results. Also, it would be very interesting to test the proposed model in other tasks involving structured information.

Acknowledgments. Work partially supported by the Spanish MEC under the STraDA research project (TIN2012-37475-C02-01), the Spanish project 2010-CONES-00029, the FPU grant (AP2009-4363), and through the EU 7th Framework Programme grant tranScriptorium (Ref: 600707).

References

[1] F. Shafait, D. Keysers, T. Breuel, Performance evaluation and benchmarking of six-page segmentation algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 941–954.

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, ICDAR 2013 Competition on Historical Newspaper Layout Analysis (HNLA 2013), in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, 2013, pp. 1454–1458.

[3] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, ICDAR 2013 Competition on Historical Book Recognition (HBR 2013), in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, 2013, pp. 1459–1463.

[4] C. An, H. Bird, P. Xiu, Iterated document content classification, in: Proc. of ICDAR, volume 1, Brazil, 2007, pp. 252–256.

[5] M. Bulacu, R. Koert, L. Schomaker, T. Zant, Layout analysis of handwritten historical documents for searching the archive of the cabinet of the Dutch queen, in: Proc. of ICDAR, volume 1, Brazil, 2007, pp. 23–26.

[6] M. Lemaitre, E. Grosicki, E. Geoffrois, F. Prêteux, Layout analysis of handwritten letters based on textural and spatial information and a 2D markovian approach, in: Proceedings 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008), Montreal, 2008, pp. 451–456.

[7] L. Likforman-Sulem, A. Zahour, B. Taconet, Text line segmentation of historical documents: a survey, International Journal of Document Analysis and Recognition 9 (2007) 123–138.

[8] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, F. Casacuberta, Integrated Handwriting Recognition and Interpretation using Finite-State Models, IJPRAI 18 (2004) 519–539.

[9] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, Historical document layout analysis competition, in: Proc. of ICDAR, Beijing, China, 2011, pp. 1516–1520.

[10] V. Romero, A. Fornés, N. Serrano, J. Sánchez, A. Toselli, V. Frinken, E. Vidal, J. Lladós, The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition, Pattern Recognition 46 (2013) 1658–1669.

[11] A. Esteve, C. Cortina, A. Cabré, Long term trends in marital age homogamy patterns: Spain, 1992-2006, Population 64 (2009) 173–202.

[12] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, An introduction to variational methods for graphical models, Machine Learning 37 (1999) 183–233.

[13] M. J. Wainwright, M. I. Jordan, Graphical Models, Exponential Families, and Variational Inference, volume 1, Now Publishers Inc., Hanover, MA, USA, 2008.

[14] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

[15] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Analysis and Machine Intelligence 23 (2001) 1222–1239.

[16] J. S. Yedidia, W. T. Freeman, Y. Weiss, Exploring artificial intelligence in the new millennium, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003, pp. 239–269.

[17] F. Cruz, O. R. Terrades, Document segmentation using relative location features, in: Proc. of ICPR, Japan, 2012, pp. 1562–1565.

[18] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, Multi-class segmentation with relative location prior, Int. Journal of Computer Vision 80 (2008) 300–316.

[19] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proc. of ICML, USA, 2001, pp. 282–289.

[20] A. Jain, A. Namboodiri, J. Subrahmonia, Structure in online documents, in: Proc. of ICDAR, volume 1, 2001, pp. 844–848.

[21] J. Handley, A. Namboodiri, R. Zanibbi, Document understanding system using stochastic context-free grammars, Proc. of ICDAR 1 (2005) 511–515.

[22] S. Crespi Reghizzi, M. Pradella, A CKY parser for picture grammars, Information Processing Letters 105 (2008) 213–217.

[23] F. Álvaro, J. Sánchez, J. Benedí, Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models, Pattern Recognition Letters 35 (2014) 58 – 67.

[24] F. Álvaro, F. Cruz, J. Sánchez, O. Ramos-Terrades, J. Benedí, Page segmentation of structured documents using 2d stochastic context-free grammars, in: 6th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), LNCS 7887, Springer, 2013, pp. 133–140.

[25] V. Romero, J. Sánchez, Category-based language models for handwriting recognition of marriage license books, in: International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 788–792.

[26] S. Z. Li, Markov Random Field Modeling in Computer Vision, Springer-Verlag, London, UK, UK, 1995.

[27] D. Roth, On the hardness of approximate reasoning, Artificial Intelligence 82 (1996) 273–302.

[28] G. F. Cooper, The computational complexity of probabilistic inference using bayesian belief networks, Artif. Intell. 42 (1990) 393–405.

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

[30] Y. Weiss, Correctness of local probability propagation in graphical models with loops, 2000.

[31] J. Besag, On the statistical analysis of dirty pictures, Journal of the Royal Statistical Society. Series B (Methodological) 48 (1986) 259–302.

[32] J. Goodman, Semiring parsing, Computational Linguistics 25 (1999) 573–605.

[33] J. Ilonen, J.-K. Kamarainen, H. Kälviäinen, Fast extraction of multi-resolution Gabor features, in: 14th International Conference on Image Analysis and Processing, Modena, Italy, 2007, pp. 481–486.

[34] I. Fogel, D. Sagi, Gabor filters as texture discriminator, Biological Cybernetics 61 (1989) 103–113.