Page Segmentation of Structured Documents Using 2D Stochastic Context-Free Grammars

Francisco Álvaro[1], Francisco Cruz[2], Joan-Andreu Sánchez[1], Oriol Ramos Terrades[2], and José-Miguel Benedí[1]

[1] Instituto Tecnológico de Informática, Universitat Politècnica de València.
{falvaro,jandreu,jbenedi}@dsic.upv.es
[2] Centre de Visió per Computador, Universitat Autònoma de Barcelona.
{fcruz,oriolrt}@cvc.uab.cat

Abstract. In this paper we define a bidimensional extension of Stochastic Context-Free Grammars for page segmentation of structured documents. Two sets of text classification features are used to perform an initial classification of each zone of the page. Then, the page segmentation is obtained as the most likely hypothesis according to a grammar. This approach is compared to Conditional Random Fields and results show significant improvements in several cases. Furthermore, grammars provide a detailed segmentation that allowed a semantic evaluation which also validates this model.

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

1 Introduction

Page segmentation is a challenging problem of document image analysis which is important for further document processing problems. The page segmentation problem consists of detecting relevant zones in a given document. The relevant zones depend on the task and they can be images, textual zones, special symbols, etc. Many successful page segmentation techniques have been defined for typeset documents. Most of the techniques are based on a first step of classification at pixel level and then the pixels are grouped into regions to obtain uniform zones [2].

In case of historical handwritten documents, the challenge in page segmentation is to detect homogeneous handwritten zones [4]. Correct detection of homogeneous textual zones is important for tackling further problems like transcription or word spotting. Many historical handwritten documents exhibit regularities like in typeset documents, and page segmentation techniques used for typeset documents can be considered for this type of documents [3]. 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 that

documents, holistic approaches that consider all elements seem more appropriate. This research is focused on this second type of historical handwritten documents, concretely in marriage license books (see Figure 1). Similar structured documents have been studied in other researches [4].

Marriage license books are documents that were used for centuries to record marriages in ecclesiastical institutions. Each record is associated to a marriage and the transcription of these documents has been considered very interesting for demography and migratory researches [7]. Each record is composed of several textual zones that compose a logical structure. Two relevant page segmentation problems can be stated for these documents. First, to classify the different textual zones of the records. And second, to isolate the logically related parts of each record of a given page.

In [6], a solution is proposed for classifying the different textual zones that are present in marriage license books. In that research, pixel classification based on texture features obtained from the Gabor transform are compared with Relative Location Features [8]. Both sort of features are combined with Conditional Random Fields [11] in order to take into account contextual information in the classification process of the pixels.

In this paper we will study the use of bidimensional Stochastic Context-Free Grammars (2D SCFG) for the two problems of page segmentation of structured documents. SCFG have been used previously for document image analysis [10, 9]. 2D SCFG is a well known formalism that has been studied in the past for bidimensional parsing [5, 1]. 2D SCFG are able to represent efficiently contextual bidimensional relations that are important for page segmentation.

In the following section we describe the problem of page segmentation applied to marriage license books. 2D SCFG are introduced in Section 3, and the corresponding parsing algorithms are also described in this section. Section 4 describes the features used for classifying textual zones at local level. Finally, Section 5 reports the experiments that were carried out.

2 Problem description

Marriage license books are handwritten documents that have been used in ecclesiastical institutions for centuries for registering marriages. Most of these documents have a similar structure to an accounting book. In these documents, the pages were very orderly written, and the structure in each page was quite rigid.

These documents are composed of pages, and each page contains several records (see Figure 1). Each record has a *husband surname's block* (Figure 1.a), the *main block* (Figure 1.b), and the *tax block* (Figure 1.c). Note that the documents can have additional textual zones, like the date that can be seen at the beginning of the page, but in this paper these additional zones will be ignored.

The problem is to isolate correctly each record in each page, and to relate its corresponding parts, that is, the surname, the body text and the tax associated to each entry. In this paper we will focus in detecting the bounding boxes around the main parts of each record. Note, that a fine-grained detection of the

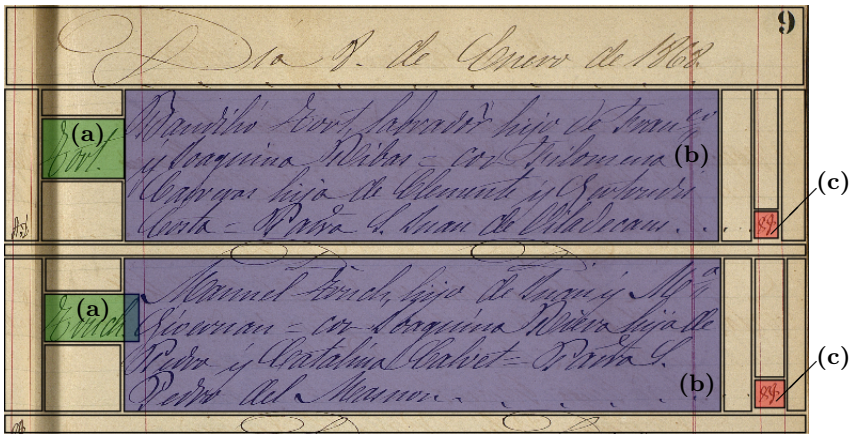

Fig. 1. Example of the page segmentation problem. Several background zones are considered and each record is composed of three parts: (a) Name (b) Body (c) Tax.

frontiers of each zone would be ideal, but this is difficult because sometimes two zones overlap (see the lower record in Figure 1). In this paper we will focus in detecting the bounding boxes around the main parts of each record. The problem of detecting the records can be stated as two different problems: first, to classify the textual zones into the previously mentioned classes; and second, to detect the complete set of records of each page and isolate each of them. This last problem is sometimes referred as *logical layout*. In the following section, 2D SCFG will be proposed to solve both problems.

3 2D Stochastic Context-Free Grammars

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 starting 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 to non-terminal symbol A. This type of grammars can be represented in Chomsky Normal Form (CNF) resulting in only two type 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. 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 must be arranged according to the spatial relation constraint.

Given a page image, the problem is to obtain the most likely segmentation 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 [5] to include the stochastic information of our model.

The CYK algorithm for 2D SCFG is based on building a parsing table $\mathcal{T}$. Each hypothesis accounts for a region $z_{x,y;x+i,y+j}$. Each region z is defined as a rectangle delimited by its top-left corner (x, y) and its bottom-right corner $(x + i, y + j)$, and $\ell_{i\times j}$ represents a subproblem of size $i \times j$. An element of the parsing table $\mathcal{T}_{x,y;x+i,y+j}[A]$ represents that nonterminal A can derive the region $z_{x,y;x+i,y+j}$ with certain probability.

The parsing process is defined as a dynamic programming algorithm. The derivation probability of a certain region is marginalized according to the terminal classes c and taking into account the size of the regions. The most likely parsing is approximated by maximizing the posterior probability and by considering usual assumptions, the parsing table is initialized for each region z of size 1×1 as

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

where $P(A \to c)$ is the probability of the model for terminal c; $P(c \mid z)$ represents the probability that region z belongs to class c and it is described in Section 4; and $P(\ell_{1\times 1} \mid A)$ is the probability that nonterminal A derives a subproblem of size 1×1. Second, the algorithm continues analyzing new hypotheses of increasing size using the binary rules of the grammar. The general case is computed for all i and j, $2 \le i \le w,\ 2 \le j \le h$ as

$$\mathcal{T}_{x,y;x+i,y+j}[A] = \\ P(\ell_{i\times j}|A) \max_{B,C} \begin{cases} \max\limits_{1<k\le i} P(A \xrightarrow{\mathrm{H}} B\,C)\, \mathcal{T}_{x,y;x+k,y+j}[B]\, \mathcal{T}_{x+k+1,y;x+i,y+j}[C] \\ \max\limits_{1<k\le j} P(A \xrightarrow{\mathrm{V}} B\,C)\, \mathcal{T}_{x,y;x+i,y+k}[B]\, \mathcal{T}_{x,y+k+1;x+i,y+j}[C] \end{cases} \tag{2}$$

where a new subproblem $\mathcal{T}_{x,y;s,t}[A]$ is computed from two smaller subproblems. The probability is maximized for every possible vertical and horizontal decomposition resulting in the region $z_{x,y;s,t}$. 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 expected size for a nonterminal $P(\ell_{i\times j} \mid A)$ as a source of information.

The expected size $P(\ell_{i\times j} \mid A)$ 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 an specific nonterminal for each zone of interest. This can be seen in Figure 1 where the expected size of the background region on top of the page will be different than the expected 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.

Finally, the most likely parsing of the full input page is obtained in $\mathcal{T}_{0,0;w,h}[S]$. 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 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 [6] we used two different set of features, Gabor features as texture descriptors and Relative Location Features [8].

4.1 Texture features

A common procedure to analyze the texture information of a document image is the use of a multi-resolution filter bank in order to compute a set of responses for several orientations and frequencies of the filter. For this research we defined the set of features $x \in \mathbb{R}^{m \times n}$ which corresponds with the set of Gabor filter responses computed for m frequencies and n orientations at pixel level.

Once we obtained the set of features, we calculated the probability $P(c \mid z)$ required in Eq. (1) by computing the probability density function for each possible class by means of a Gaussian mixture model (GMM).

4.2 Relative Location Features

Relative Location Features (RLF) were introduced by Gould *et al.* in [8] as a way to encode inter-class spatial relationships as local features. Each of these features, $v_c(z)$, 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.

The encoding process of the RLF comes from the union of several previous steps. Firstly, we need to compute a relative location probability map for each pair of classes. This mechanism allows us to represent the location where is most likely to find elements of one class respect to the others. Secondly, we use the information provided by the GMM previously calculated to compute an initial labeling configuration for each region. Finally, combining the information from the previous steps we obtain the final RLF configuration. Further details about the process can be seen in [6]. Once we have computed the set of RLF, we are able to compute the probability required by the 2D SCFG $P(c \mid z)$ as:

$$P(c \mid z) = w^{app} \log P_G(c \mid z) + w_c^{other} \log v_c^{other}(z) + w_c^{self} \log v_c^{self}(z) \qquad (3)$$

where w_c^{other} and w_c^{self} are the different sets of RLF and w^{app}, w_c^{other} and w_c^{self} are the corresponding weights learned from a logistic regression model.

5 Experiments

This section describes the experiments that were carried out to evaluate the proposed 2D SCFG model for page segmentation. First, we describe the used database and the experimentation carried out. Then, we report several performance metrics and results are compared to other approaches based on CRF [6].

5.1 Database

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. This paper is focused on the segmentation of volume 208 of this collection (see Figure 1 for a sample). This volume has 593 pages of which we labeled at pixel level the first 80 pages of the volume. We used the first 60 for train and the remaining 20 for test.

5.2 2D SCFG Experiments

We used a 2D SCFG to tackle the page segmentation problem on the 5CofM dataset. As we dealt with well structured documents, we manually defined the grammar to account for them. According to the model described in Section 3, we had to train several probability distributions. In this paper, these distributions were trained in a simple way. First, the text classification features were trained from the ground-truth information of the training set using cells of 50×50 pixels to obtain results comparable to those in [6]. Afterwards, the grammar production probabilities $P(A \rightarrow \alpha)$ were learnt by using Viterbi estimation. Likewise, the expected sizes $w \times h$ for each nonterminal $P(\ell_{w \times h} \mid A)$ were estimated by considering the number of occurrences in the training set.

The training set had 60 pages and several stochastic distributions were involved in the parsing process. In order to make the system able to account for unseen events and to obtain a better balance among the probability distributions, we smoothed both the production probabilities and the expected sizes of nonterminals. We simply smoothed the production probabilities by setting a minimum probability threshold, and we applied a 2D Gaussian filter to smooth the expected sizes of nonterminals.

In order to obtain the best smoothing parameters we performed a cross-validation experiment. The training set was split in 6 blocks each one comprising 10 pages. Several experiments were conducted considering 5 blocks for training and the remaining one for validation. Finally, we classified the test set by using the model learnt from the training set and the best parameters of the validation experimentation. Table 1 shows the segmentation results for each class. The reported metrics are the average precision, recall and F-measure computed at $(d \times d)$-cell level for each page in the test set.

Results showd that classes *Body* and *Name* are classified with good F-measure rates, whereas class *Tax* represents the most challenging part. It is

Table 1. Page segmentation results for different models and text classification features.

Model	Class	Gabor			Gabor + RLF		
		Precision	Recall	F-measure	Precision	Recall	F-measure
CRF	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
2D SCFG	Body	0.88	0.97	0.92	0.88	0.97	0.92
	Name	0.72	0.93	0.81	0.82	0.83	0.82
	Tax	0.39	0.94	0.54	0.63	0.69	0.64

related with the percentage of the page that represents each region, because it is more difficult to properly classify small regions.

There are two factors to take into account: the text classification features and the page segmentation model. Regarding text classification features, results using CRF as page segmentation model showed that the inclusion of the RLF information significantly outperformed the results obtained with Gabor features. However, RLF did not produce such a great improvement when grammars were the segmentation model. 2D SCFG took advantage of the knowledge about the document structure, thus, grammars were able to overcome the lacks of Gabor features obtaining good results for the classes *Body* and *Name* without the additional spatial information provided by RLF. Even though, the inclusion of RLF in 2D SCFG classification slightly improved the segmentation results for class *Name*, as well as it achieved a good f-measure value for class *Tax*.

Comparing the performance of 2D SCFG with CRF, results showed that grammars achieved a great improvement when Gabor features were used, which was not so prominent with RLF features. To assert whether this improvement was statistically significant or not we computed a two-sided Wilcoxon rank sum test with a significance level of 5%. This test proved that in the case of Gabor features, 2D SCFG significantly improved the F-measure rates for classes *Body* and *Name*, whereas the enhancement for the *Tax* class was not statistically significant with a p-value of 0.39. Otherwise, the use of RLF only improved the results on the *Name* class in comparison to the CRF results.

CRF recognized regions but did not identify the explicit segmentation in records. However, 2D SCFG can provide the coordinates of every region in the most likely hypothesis. Using this information we also carried out a semantic evaluation. We manually evaluated the results attending to two different metrics: the number of records properly segmented in terms of F-measure; and the numbers of lines correctly detected into the *Body* of the proper record in views of a future line segmentation process. Results showed that the use of RLF improved the final records detection as well as the number of lines properly segmented with F-measure values of 0.80 and 0.89 respectively. These results also validated the proposed model for page segmentation given that the records detection is the final goal of this task and 2D SCFG obtained good results.

6 Conclusions and future work

In this paper we have proposed a model based on 2D SCFG for page segmentation of structured documents. Two sets of features have been tested and both provided good results for this task, where RLF obtained better results. The experimentation carried out proved that this approach improved the results obtained by previous works with CRF. Furthermore, grammars were able to provide the detailed information of the page segmentation, hence, semantic evaluation could be done and results showed a good performance of the model.

Future work will be focused in testing other text classification features that could improve the segmentation results. Moreover, we expect that results will be enhanced using a smaller cell size, specially for class *Tax* and the overlapping area between regions *Name* and *Body*.

Acknowledgments. Work supported by the EC (FEDER/ FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the MITTRAL (TIN2009-14633-C03-01) project, the FPU grant (AP2009-4363), and by the Generalitat Valenciana under the grant Prometeo/2009/014.

References

1. F. Álvaro, J.A. Sánchez, and J.M. Benedí. Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models. *Pattern Recognition Letters*, 2012.
2. C. An, H.S. Bird, and P. Xiu. Iterated document content classification. In *Proc. of ICDAR*, volume 1, pages 252–256, Brazil, 2007.
3. A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. Historical document layout analysis competition. *Proc. of ICDAR*, pages 1516–1520, 2011.
4. M. Bulacu, R. Koert, L. Schomaker, and 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, pages 23–26, Brazil, 2007.
5. S. Crespi Reghizzi and M. Pradella. A CKY parser for picture grammars. *Information Processing Letters*, 105(6):213–217, February 2008.
6. F. Cruz and O. Ramos Terrades. Document segmentation using relative location features. In *Proc. of ICPR*, pages 1562–1565, Japan, 2012.
7. A. Esteve, C. Cortina, and A. Cabré. Long term trends in marital age homogamy patterns: Spain, 1992-2006. *Population*, 64(1):173–202, 2009.
8. S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *Int. Journal of Computer Vision*, 80(3):300–316, 2008.
9. J.C. Handley, A.M. Namboodiri, and R. Zanibbi. Document understanding system using stochastic context-free grammars. *Proc. of ICDAR*, 1:511–515, 2005.
10. A.K. Jain, A.M. Namboodiri, and J. Subrahmonia. Structure in online documents. In *Proc. of ICDAR*, volume 1, pages 844–848, 2001.
11. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289, USA, 2001.