Neurocomputing 150 (2015) 147–154

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# 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, Spain
[b] Centre de Visió per Computador, Universitat Autònoma de Barcelona, Spain

ARTICLE INFO

*Article history:*
Received 31 January 2014
Received in revised form
2 July 2014
Accepted 9 August 2014
Available online 22 October 2014

*Keywords:*
Document image analysis
Stochastic context-free grammars
Text classification features

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.



## 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 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 This paper is focused on determining the structure and the segmentation of textual zones in images of handwritten historical documents.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]. Successful 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 Fig. 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

* Corresponding author.
*E-mail addresses:* falvaro@prhlt.upv.es (F. Álvaro), fcruz@cvc.uab.cat (F. Cruz), jandreu@prhlt.upv.es (J.-A. Sánchez), oriolrt@cvc.uab.cat (O. Ramos Terrades), jmbenedi@prhlt.upv.es (J.-M. Benedí).

http://dx.doi.org/10.1016/j.neucom.2014.08.076
0925-2312/

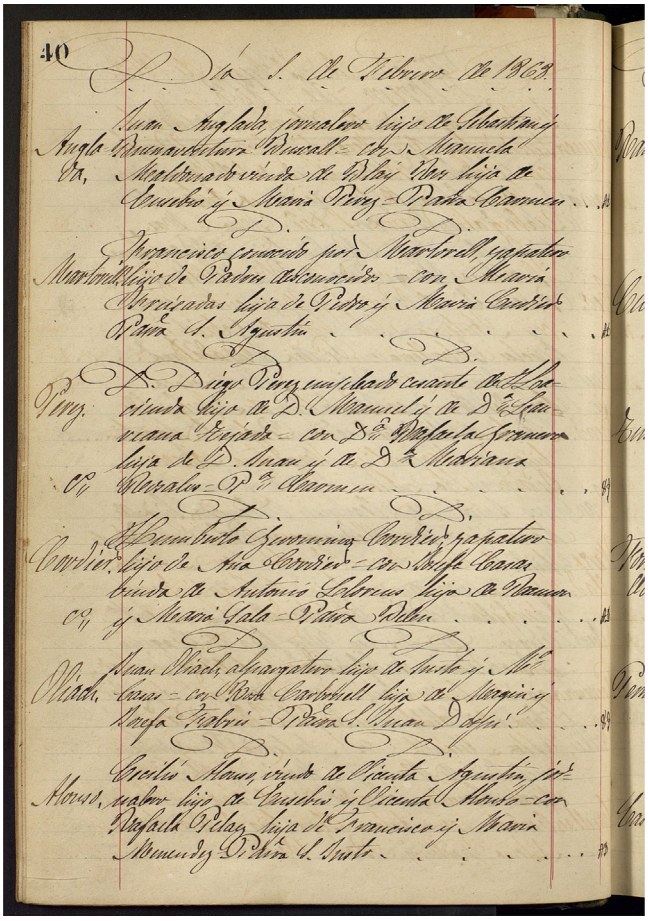

**Fig. 1.** Example of page of a marriage license book containing six records.

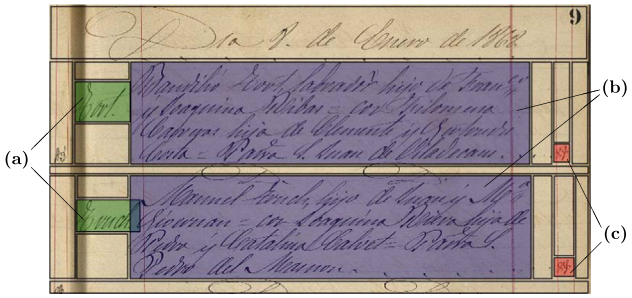


**Fig. 2.** Example of the page segmentation problem for two records. Several background zones are considered and each record is composed of three parts: (a) Name; (b) body; (c) tax.

to represent dependencies between them [12–14]. 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 is 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. Fig. 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* (Fig. 2a), the *main block* (Fig. 2b), and the *tax block* (Fig. 2c). 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

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